Cena 13,00 zł
(VAT 8%)

Indeks 381306
e-ISSN 2543-8476
PL ISSN 0043-518X

WIADOMOŚCI STATYSTYCZNE

THE POLISH STATISTICIAN

LUTY / FEBRUARY
ROK / VOLUME 67

2022 | 2

GŁÓWNY URZĄD STATYSTYCZNY
STATISTICS POLAND

POLSKIE TOWARZYSTWO STATYSTYCZNE
POLISH STATISTICAL ASSOCIATION

WIADOMOŚCI STATYSTYCZNE

THE POLISH STATISTICIAN

LUTY / FEBRUARY
ROK / VOLUME 67

2022 | 2 (729)

ADRES REDAKCJI / EDITORIAL OFFICE ADDRESS

Główny Urząd Statystyczny / Statistics Poland, al. Niepodległości 208, 00-925 Warszawa
tel./phone +48 22 608 32 25, e-mail: redakcja.ws@stat.gov.pl

Redakcja językowa: Wydział Czasopism Naukowych, Główny Urząd Statystyczny
Language editing: Scientific Journals Division, Statistics Poland

Redakcja techniczna, skład i łamanie, opracowanie materiałów graficznych, korekta: Zakład Wydawnictw Statystycznych – zespół pod kierunkiem Wojciecha Szuchty
Technical editing, typesetting, preparation of graphic materials, proofreading: Statistical Publishing Establishment – team supervised by Wojciech Szuchta

Zakład Wydawnictw Statystycznych

Druk i oprawa / Printed and bound:
Zakład Wydawnictw Statystycznych / Statistical Publishing Establishment
al. Niepodległości 208, 00-925 Warszawa, zws.stat.gov.pl

Wersja elektroniczna, stanowiąca wersję pierwotną czasopisma, jest dostępna na ws.stat.gov.pl
The primary version of the journal is issued in electronic form, available at ws.stat.gov.pl

Indeks 381306

Informacje w sprawie sprzedaży czasopisma / Sales of the journal:
Zakład Wydawnictw Statystycznych / Statistical Publishing Establishment
tel./phone +48 22 608 32 10, +48 22 608 38 10
Prenumerata jest prowadzona przez / Subscription is available at RUCH S.A.
Zamówienia na prenumeratę można składać na stronie / Subscriptions can be ordered at
www.prenumerata.ruch.com.pl

Zam. 55/2022 – nakład 300 egz.

SPIS TREŚCI
CONTENTS

OD REDAKCJI

Lutowy numer „Wiadomości Statystycznych. The Polish Statistician" zawiera artykuły dotyczące zastosowań statystyki w praktyce, edukacji statystycznej i historii statystyki, a także sprawozdanie z konferencji.

Mgr Agata Warchalska-Troll i dr Tomasz Warchalski w artykule *The selection of areas for case study research in socio-economic geography with the application of* k*-means clustering* badają możliwości zastosowania powszechnie używanej w analizie skupień metody *k*-średnich do wyboru jednostek przestrzennych (w tym przypadku gmin) w celu przeprowadzenia studiów przypadku. Na przykładzie pogłębionych badań nad relacją między ochroną przyrody a rozwojem lokalnym i regionalnym w polskich Karpatach pokazują, jak można rozwiązać problem metodyczny polegający na optymalnym wyznaczeniu gmin. Z przeprowadzonych analiz (wykorzystujących m.in. optymalizację liczby skupień za pomocą metody łokcia i statystyki pseudo-*F*) wynika, że metoda *k*-średnich, pomimo pewnych słabości, może być przydatna do tworzenia klasyfikacji i typologii służących do wyboru obszarów do studiów przypadku ze względu na jej użyteczność oraz dostępność w oprogramowaniu typu *open source*. Autorzy zauważają jednak, że jej zastosowanie może wymagać wsparcia poprzez analizę dodatkowych źródeł informacji lub skorzystanie z wiedzy eksperckiej.

Przedmiotem pracy dr. Stanisława Jaworskiego *O różnych kryteriach konstrukcji przedziałów ufności na przykładzie rozkładu normalnego* są kryteria służące konstruowaniu przedziałów ufności oparte na kontroli błędu przeszacowania i niedoszacowania parametru, jak również na zadanej arbitralnie precyzji oszacowania tego parametru. Autor prezentuje zasady i narzędzia konstrukcji przedziałów ufności, a szczególną uwagę poświęca sytuacji, gdy istnieje potrzeba zastosowania niesymetrycznego podziału ryzyka błędu. Zamieszczone w pracy przykłady pokazują wpływ zadanych kryteriów na postać przedziałów ufności, a wykonane obliczenia przedstawiają zależność szerokości danego przedziału od rozmiaru próby i odwrotnie – rozmiaru próby od zadanej szerokości tego przedziału. Artykuł uwidacznia, jak można rozszerzyć temat konstrukcji przedziałów ufności w edukacji statystycznej.

The 1897 Census in the Kingdom of Poland autorstwa dr. hab. Piotra Rachwała traktuje o zagadnieniach związanych z organizacją, przebiegiem i opracowaniem wyników pierwszego powszechnego spisu ludności w Królestwie Polskim, który przeprowadzono w 1897 r. Autor ukazuje problematykę spisową w kontekście działalności władz centralnych w Petersburgu. Podkreśla wykorzystanie nowoczesnych technologii, przede wszystkim maszyn liczących Holleritha i kodowania informacji na kartach perforowanych, co znacząco usprawniło obliczenia. Autor zaznacza, że przyjęcie krytycznego podejścia w badaniu danych uzyskanych w wyniku tego spisu (wiarygodność części z nich może budzić zastrzeżenia) pozwala na efektywne wykorzystanie bogactwa zgromadzonej wówczas wiedzy.

W bieżącym numerze znajdą Państwo również – opracowane przez dr hab. Grażynę Dehnel, prof. UEP, dr. hab. Andrzeja Młodaka, prof. AK, dr. hab. Tomasza Klimanka, prof. UEP, i dr. hab. Marcina Szymkowiaka, prof. UEP – sprawozdanie z międzynarodowej konferencji naukowej *Joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality*, która odbyła się w dniach 1–3 grudnia 2021 r. na Uniwersytecie Ekonomicznym w Poznaniu. To cykliczne wydarzenie, które tym razem zorganizowano w naszym kraju, stanowiło ważne forum dyskusji najwybitniejszych światowych specjalistów z zakresu kontroli ujawniania danych na temat doskonalenia metod i narzędzi ochrony informacji wrażliwych przed ich powiązaniem z konkretną jednostką (np. osobą czy podmiotem gospodarczym) oraz wypracowania równowagi między minimalizacją ryzyka dotyczącego tego powiązania a istotną dla użytkownika minimalizacją straty informacji spowodowanej podjętymi działaniami ochronnymi.

Wydanie zamyka prezentacja nowości wydawniczych GUS przygotowana przez Justynę Gustyn.

Życzymy miłej lektury.

FROM THE EDITORIAL TEAM

The February issue of *Wiadomości Statystyczne. The Polish Statistician* presents articles on the practical applications of statistics, statistical education and the history of statistics, as well as a report on a conference.

In their article *The selection of areas for case study research in socio-economic geography with the application of* k*-means clustering*, Agata Warchalska-Troll, MSc, and Tomasz Warchalski, PhD, examine the possibilities of applying the *k*-means method (commonly used in cluster analysis) for the selection of spatial units (here: gminas) for case studies. Based on in-depth research into the relationship between environmental protection and local and regional development in the Polish Carpathians, the authors demonstrate the process of solving a methodical problem involving the optimal selection of gminas. The performed analyses (which also included the optimisation of the number of clusters by means of the elbow method and the pseudo-*F* statistic) show that the *k*-means method, despite its certain shortcomings, may prove useful in the construction of classifications and typologies facilitating the choice of areas for case studies. The technique is additionally available in open source software. The authors notice, however, that the *k*-means method may need to be supplemented by additional data source analysis and expert knowledge.

Dr Stanisław Jaworski's article entitled *On various criteria for the construction of confidence intervals on the example of normal distribution* focuses on the criteria of constructing confidence intervals based on the control of the error of the overestimation and underestimation of a parameter, as well as on an arbitrarily predetermined precision of the parameter estimation. The author presents the principles and tools essential for the construction of confidence intervals, and devotes particular attention to those circumstances which require the application of an asymmetric distribution of the error risk. The examples which the study provides illustrate the influence of the predetermined criteria on the formula of confidence intervals, while the performed calculations show how the width of a given interval depends on the sample size and vice versa – how the sample size depends on the given width of this interval. The paper demonstrates how the issue of confidence interval construction may be extended in statistical education.

The 1897 Census in the Kingdom of Poland by Piotr Rachwał, PhD, DSc, presents issues concerning the organisation, course and processing of the results of the first census held in the Kingdom of Poland in 1897. The author addresses the census-related matters in the context of the activity of the central authorities in St. Petersburg. He emphasises the application of modern technologies, particularly Hollerith's calculating machines and the coding of information on perforated cards, which greatly facilitated the calculations. The author underlines that assuming a critical approach in the research of the data accumulated during the census (as the reliability of some of them are questionable) allows the effective use of the vast knowledge acquired from that survey.

The current issue also presents a report on the *Joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality* international conference held at Poznań University of Economics and Business (PUEB) on 1–3 December, 2020. The report was prepared by Grażyna Dehnel, PhD, DSc, professor at PUEB, Andrzej Młodak, PhD, DSc, professor at the Calisia University, Tomasz Klimanek, PhD, DSc, professor at PUEB, and Marcin Szymkowiak, PhD, DSc, professor at PUEB. This cyclical conference, this time held in Poland, provided a significant forum for the discussion of the world's most prominent specialists in the field of data disclosure control. The meeting focused on the improvement of the tools and methods of protecting sensitive information from its connection to a particular entity (an individual or business) and on creating a balance between the minimisation of the risk relating to this connection and the minimisation of the loss of the information significant to the user, resulting from the undertaken protection activity.

The issue concludes with Justyna Gustyn's presentation of Statistics Poland's new publications.

We wish you pleasant reading.

Wiadomości Statystyczne. The Polish Statistician, 2022, vol. 67, 2, 1–20 DOI: 10.5604/01.3001.0015.7717
Statystyka w praktyce / Statistics in practice
Zgłoszony/submitted: 1.08.2021, poprawiony/revised: 9.01.2022, zaakceptowany/accepted: 4.02.2022

The selection of areas for case study research in socio-economic geography with the application of *k*-means clustering

Agata Warchalska-Troll,[a] Tomasz Warchalski[b]

Abstract. The grouping techniques which are known in statistics are rarely used by geographers to select a research area. The aim of the paper is to examine the potential use of the *k*-means clustering (partitioning) method for the selection of spatial units (here: gminas, i.e. the lowest administrative units in Poland) for case studies in socio-economic geography. We explored this topic by solving a practical problem consisting in the optimal designation of gminas for in-depth research on the interaction between nature protection and local and regional development in the Polish Carpathians. Particular attention was devoted to defining an appropriate number of clusters by means of the elbow method as well as the pseudo-*F* statistic (the Calinski-Harabasz index). The data for the analysis were mostly provided by Statistics Poland and covered the period of 1999–2012. The multi-stage procedure resulted in the selection of the following gminas: Cisna, Lipinki, Ochotnica Dolna, Sękowa, Szczawnica and Zawoja.

The example described in the paper demonstrates that the *k*-means technique, despite its certain deficiencies, may prove useful for creating classifications and typologies leading to the selection of case study sites, as it is relatively time-effective, intuitive and available in open-source software. At the same time, due to the complexity of the socio-economic characteristics of the areas, the application of this method in socio-economic geography may require support in terms of the interpretation of the results through the analysis of additional data sources and expert knowledge.

Keywords: case study, *k*-means partitioning, elbow method, pseudo-*F* statistic, Calinski-Harabasz index

JEL: C38, O18, R58

Wybór obszarów do studiów przypadku w geografii społeczno-ekonomicznej z zastosowaniem metody grupowania *k*-średnich

Streszczenie. Znane w statystyce techniki grupowania są rzadko wykorzystywane przez geografów do wyboru obszaru badań. Celem analiz opisanych w artykule było sprawdzenie możliwości zastosowania metody podziału *k*-średnich do wyboru jednostek przestrzennych

[a] Instytut Rozwoju Miast i Regionów, Warszawa, Polska; Uniwersytet Jagielloński w Krakowie, Instytut Geografii i Gospodarki Przestrzennej, Polska / Institute of Urban and Regional Development, Warsaw, Poland; Jagiellonian University in Krakow, Institute of Geography and Spatial Management, Poland. ORCID: https://orcid.org/0000-0003-1314-3206. Autor korespondencyjny / Corresponding author, e-mail: agata.warchalska.troll@gmail.com.

[b] Badacz niezależny, Polska / Independent researcher, Poland. ORCID: https://orcid.org/0000-0002-2894-2265. E-mail: twarchalski@gmail.com.

(w tym przypadku gmin) do studiów przypadku. Dokonano tego poprzez rozwiązanie problemu metodycznego polegającego na optymalnym wyznaczeniu gmin do pogłębionych badań nad relacją między ochroną przyrody a rozwojem lokalnym i regionalnym w polskich Karpatach. Szczególną uwagę zwrócono na określenie odpowiedniej liczby skupień za pomocą metody łokcia (ang. *elbow method*) oraz statystyki pseudo-*F* (wskaźnika Calińskiego-Harabasza). Dane wykorzystane w analizach pochodziły z Głównego Urzędu Statystycznego i obejmowały okres 1999–2012. W rezultacie kilkustopniowej procedury wytypowano gminy: Cisna, Lipinki, Ochotnica Dolna, Sękowa, Szczawnica i Zawoja.

Opisany w artykule przykład pokazuje, że metoda *k*-średnich, pomimo pewnych słabości, może być przydatna do tworzenia klasyfikacji i typologii prowadzących do wyboru obszarów do studiów przypadku ze względu na jej użyteczność oraz dostępność w oprogramowaniu typu *open source*. Zarazem jednak – z uwagi na stopień złożoności społeczno-ekonomicznych cech obszarów – zastosowanie tej metody w geografii społeczno-ekonomicznej może wymagać wsparcia interpretacji jej wyników analizą dodatkowych źródeł informacji oraz wiedzą ekspercką.

Słowa kluczowe: studium przypadku, grupowanie metodą *k*-średnich, metoda łokcia, statystyka pseudo-*F*, wskaźnik Calińskiego-Harabasza

1. Introduction

The selection of a study area is a fundamental stage in geographic research, regardless of the specialisation involved. By its very nature, research in geography is often based on case studies of areas selected for their uniqueness or, in contrast, for being typical in terms of certain features. This basic dichotomy of the nature of the case study as a research strategy underlies its various classifications (see for instance Taylor, 2016, p. 583). In most cases the choice of a study unit(s) needs to be presented against a broader spatial background and sometimes additionally justified by a more detailed comparative analysis. The latter gains particular importance when the line of argument is based on the representativeness of the studied area.

This paper addresses the practical issues relating to the statistical techniques applied to official statistical data. The aim of the paper is to examine the potential use of the *k*-means partitioning method for the selection of spatial units (here: gminas, i.e. the lowest administrative units in Poland) for case studies in socio-economic geography. In its substantial part, the analyses presented in this paper were prepared within a research project investigating selected aspects of the influence of national parks and NATURA 2000 sites on local and regional development in the Polish Carpathians.[1] The study focused on three aspects: (1) technical infrastructure, (2) economic activity and its structure, and (3) social development. The project was initiated in 2013, with its fundamental part completed in 2019, and consisted of stages including desk and field research. When the research reached its summarising phase, it was noted that *k*-means clustering (partitioning) is still rather rarely used in socio-economic geography in the procedure of selecting the case study area (for more details please see the overview presented later in this

[1] Some results of this project were already published (Warchalska-Troll, 2018, 2019).

section). Considering the above, the authors decided that the applied research scheme is worth exploring, thus the paper focuses on discussing its strengths and weaknesses on the basis of actual research carried out in socio-economic geography.

As we reviewed the 2015–2019 issues of four Polish geographical journals of a well-established position at a national level and a growing share of foreign contributions (*Bulletin of Geography. Socio-economic series*, *Geographia Polonica*, *Polish Geographical Review*, *Regional and Local Studies*), we found that in only three cases statistical techniques were used in the process of selecting the study areas in socio-economic geography. In one case, *k*-means partitioning was applied for this purpose (Mikuš et al., 2016), although not by the authors themselves, but through a reference to another publication which was not available to us (Šebová, L. (2013). *Identifikácia marginálnych regiónov na Slovensku* [Doctoral disertation]. Comenius University in Bratislava).

From a total of 475 papers on socio-economic geography published in the above-mentioned journals, 181 included case studies. The vast majority of these articles (150 out of 181, i.e. 83%) included only a descriptive (although sometimes not even a comparative) comment on how and why a particular area(s) was selected for the case study analysis. Out of the 181 papers, 22 (i.e. 12%) provided no justification for the choice of the studied area and in the six remaining cases these justifications were reduced to one or two general phrases.[2]

Although the paper focuses on Poland, we also investigated how often '*k*-means' and 'case study' appear together in the scientific content of two high-impact databases:[3] Taylor & Francis online and Science Direct (by Elsevier). To remain discipline-specific, we narrowed our search to journals on socio-economic geography. In the Taylor & Francis online database, we found only 43 records meeting the criteria mentioned above in the field of 'Geography', and out of these 26 were in the 'Human Geography' section (which included transport geography, social geography, economic geography, regional geography and more). In none of these 26 papers *k*-means was applied in the selection of the studied sites. In the Science Direct database, our inquiry provided 914 research papers published in journals belonging to the social sciences discipline, out of which we verified the content of three journals which are often chosen by geographers and which had the highest number of items responding to the above-mentioned query: *Applied Geography* (28 results), *Landscape and Urban Planning* (26) and *Land Use Policy* (23). Here again, we found no example of the *k*-means technique used for the designation of the studied areas. Clustering was evidently broadly applied for classifications, typologies and regionalisations, but not

[2] However, it is worth noting that in numerous cases applying statistical techniques to study site selection is unnecessary or even useless, as many socio-economic geographers investigate very 'soft', qualitative matters, unique places or rarely represented features. Moreover, the understanding of case study as a research approach may vary among scientists (Babbie, 2007; Taylor, 2016). Finally, it should be taken into account that the research papers usually present only a part or a summary of broader research projects, which in consequence may also relate to the description of the applied methods.

[3] The query was executed in mid-December 2021.

in the study site selection procedure itself. This query is for sure limited in scope and gives only an idea of how popular the *k*-means technique is in socio-economic geography. However, the answer to the query at least suggests that the *k*-means and similar clustering techniques[4] are not a 'must-have' in human /socio-economic geographers' toolkit when it comes to selecting sites for their case studies.

2. Research method

2.1. Data and study area

The fields of interest described in the previous part of this paper implied our choice of variables from within the data provided by the Local Data Bank of Statistics Poland. The initial database covered 79 socio-economic variables downloaded directly from the Statistics Poland website, and further 44 variables were our own calculations based on the official statistics data. Four supplementary variables such as the list of spa cities/towns, the list of Natura 2000 sites, etc. were also taken from official governmental sources. All in all, the total database covered 127 variables (features), many of which were overlapping in terms of their substantive content. With the purpose of narrowing the set of the variables to be covered in our detailed analyses, we experimented with different subsets of 'feature-representatives' in infrastructure, economic activity, social development and environment. We then investigated the correlation (*r* Pearson) between the variables and evaluated the outcome using Student's *t*-test at a 0.05 significance level. At this point, it should be mentioned that the fact of a strong correlation existing between the variables posed no technical problem in the *k*-means analysis, which was at the core of our procedure; however, we did prefer to avoid strong correlations whenever possible. Despite this investigation and given the complexity of the research matter, the final choice of the variables could not be made without an expert decision and included: (1) the total number of tourist accommodation establishments; (2) the occupancy of bed places; (3) bed places per 1,000 citizens; (4) the share of entities in manufacturing (section C, divisions 10, 11, 15 and 16 of NACE Rev. 1.2 / PKD 2007) in the total number of the entities in the REGON register of economic units; (5) the share of entities in retail trade (except for sales of motor vehicles and motorcycles; section G, division 47 of NACE Rev. 1.2 / PKD 2007) in the total number of the entities in the REGON register; (6) the share of entities in selected services[5] (sections/divisions I56, M72, 74, P85, R90, 91, 93 and S94 of NACE Rev. 1.2 / PKD 2007) in the total number of the entities in the REGON register; (7) the share of

[4] The searching algorithm sorted the outcome by how well (thus not always in 100%) it matched the inquiry.

[5] Food and beverage service activities; scientific research and development; other professional, scientific and technical activities; education, creative, arts and entertainment activities; libraries, archives, museums and other cultural activities; sports activities; activities of membership-based organisations.

persons using the water supply system; (8) the share of persons using the sewage system; (9) outlays on fixed assets serving environmental protection and water management; (10) non-profit organisations and associations per 10,000 citizens; (11) natural persons conducting economic activity per 100 citizens; (12) entities newly included in the REGON register per 10,000 citizens (Table 1).[6]

Table 1. Basic characteristics of the variables chosen to create a classification of the Polish Carpathian gminas with the aim to select case study sites

Variable (description and unit)	Years	Value		
		min	max	mean
Total tourist accommodation establishments	2012	0.00	307.00	8.27
Occupancy of bed places in tourist accommodation establishments (total number of visitors in a given year per number of bed places in tourist accommodation establishments)[a]		0.00	158.07	16.97
Bed places in tourist accommodation establishments per 1,000 citizens[a]		0.00	825.79	44.82
Share (in %) of entities operating in manufacturing (section C, divisions 10, 11, 15 and 16 of NACE Rev. 1.2 / PKD 2007) in the total number of entities included in the REGON register[a]		0.01	0.33	0.06
Share (in %) of entities operating in retail trade except for the sales of motor vehicles and motorcycles (section G, division 47 of NACE Rev. 1.2 / PKD 2007) in the total number of entities included in the REGON register[a]		0.05	0.27	0.16
Share (in %) of entities operating in services (sections/divisions I56, M72, 74, P85, R90, 91, 93 and S94 of NACE Rev. 1.2 / PKD 2007) in the total number of entities included in the REGON register (selected services: food and beverage service activities; scientific research and development; other professional, scientific and technical activities; education, creative, arts and entertainment activities; libraries, archives, museums and other cultural activities; sports activities; activities of membership-based organisations)[a]		0.05	0.24	0.13
Entities newly added to the REGON register per 10,000 citizens (the total for the whole available period, i.e. 2009–2012)[a]	2009–2012	114.00	990.00	302.30

a Authors' calculations.

[6] In contrast to e.g. Kraszewska (2016), who investigated the regional differentiation of the risk of poverty occurring in Poland using the hierarchical Ward clustering method, in this study the variables cannot be classified as 'enhancers/boosters' and 'dampers', since the analysis aims at classifying gminas according to a selection of features and not at investigating the influence of these features on any phenomenon.

Table 1. Basic characteristics of the variables chosen to create a classification of the Polish Carpathian gminas with the aim to select case study sites (cont.)

Variable (description and unit)	Years	Value		
		min	max	mean
Natural persons conducting economic activity per 100 citizens	2012	4.10	25.40	9.68
Non-profit organisations and associations per 10,000 citizens		7.00	93.00	26.84
Persons using the water supply system in % of the total population		0.00	97.70	42.31
Persons using the sewage system in % of the total population		0.00	1.00	0.43
Outlays on fixed assets serving environmental protection and water management (the total for the whole available period, i.e. 1999–2008), *per capita* (average population for the period 1999–2008) in thousand PLN[a]	1999–2008	0.00	12.94	1.18

a Authors' calculations.

Source: authors' work based on data from the Local Data Bank of Statistics Poland.

After verifying that the variables to be used for the calculations have an acceptable coefficient of variation (in all cases it was above 20%) a min-max normalisation was performed using the following formula (Larose & Larose, 2014, pp. 26–27):

$$X^*(i) = \frac{X\ (i) - \min\ (X)}{max\ (X) - \min\ (X)},$$

where:

$X(i)$ – the value of the *ith* observation of variable X,

$X^*(i)$ – the normalised value of the *ith* observation of variable X,

$\min(X)$ – the minimum value of variable X,

$\max\ (X)$ – the maximum value of variable X.

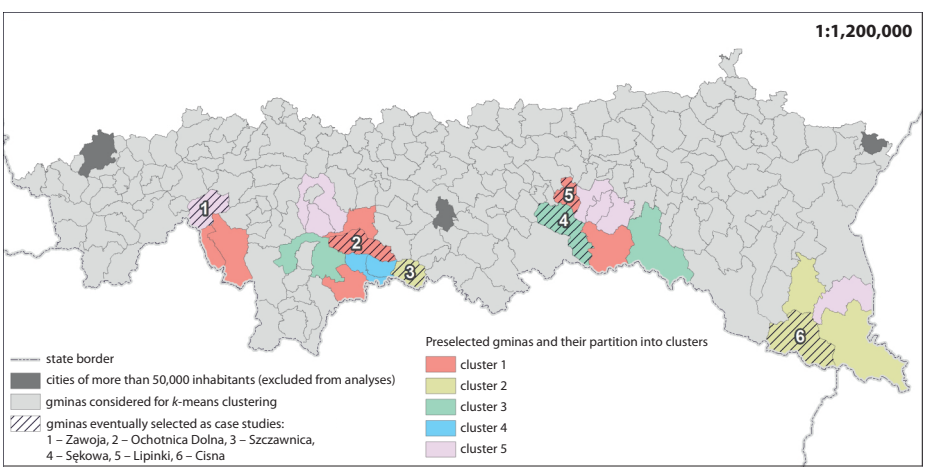
The basic analyses of data were performed in 2013 and early 2014, preceding the field-based part of the research project which was mostly carried out between 2014 and 2015. In the later stages of the case study selection procedure, several additional sources of information were used, i.a. official records (protocols) of public consultation meetings concerning the Natura 2000 sites, as well as an examination of the local press concerning gminas that were within the scope of our interest.

As regards the study area, the criterion of a gmina being listed under the Carpathian Convention (as of late 2013, when the study project was prepared) was adopted. After excluding 3 towns/cities exceeding 50,000 inhabitants, the first stage of our analysis covered 192 gminas (Figure 1).

2.2. Case study selection procedure

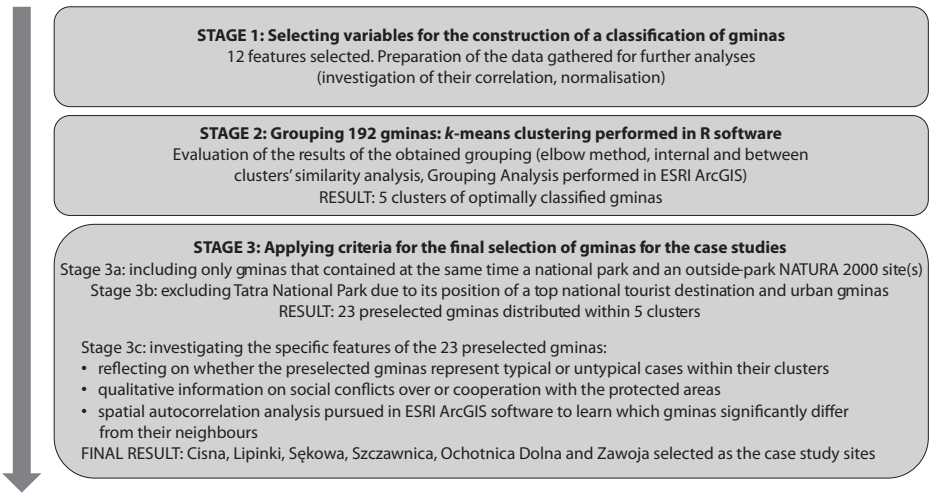
The case study selection procedure could be divided into three main stages (Figure 2). Experimenting with different variants of the clustering of a highly heterogeneous group of gminas located in the Polish Carpathians was the key part of the procedure. Once the final number of classes was chosen, we proceeded to select gminas representing each group and characterised by some additional features. This matter involved preselecting 23 gminas for closer investigation before the final six (Figure 1) were chosen, which was to guarantee that the topic of the research was covered to the greatest possible extent, and also that the number of cases to be investigated through in-depth field studies is feasible, given the limited time and budget.

Figure 1. Study area: gminas of the Polish Carpathians and the spatial scope of the analysis at different stages of the case study choice procedure



Note. Gminas selected for the case studies against the background of the whole set of Polish gminas covered by the Carpathian Convention as of late 2013 (N=195).
Source: authors' work based on the list of the Carpathian Convention gminas as of late 2013 and authors' analysis using data from the Local Data Bank of Statistics Poland; basemap: National Register of Boundaries (data of the National Geodetic and Cartographic Resource (Pol. Państwowy Zasób Geodezyjny i Kartograficzny).

Figure 2. A general framework of the case study selection procedure



Source: authors' work.

2.3. *k*-means clustering used for grouping gminas

The grouping was performed on a set of 192 gminas (defined in the previous sections of the article), using a non-hierarchical grouping method in the form of the *k*-means algorithm. This is one of the most widespread and well-established techniques whose core part was developed from the late 1950s to the late 1970s (Hartigan & Wong, 1979; Lloyd, 1982; MacQueen, 1967; Steinhaus, 1957[7]). This technique is applied in the field of cluster analysis – a branch of data analysis which involves dividing a particular set of observations into a given number of homogeneous subsets called clusters or groups. The synthetic measure of the level of the clusters' homogeneity, called the 'within groups sum of squared error' (R Core Team, n.d.) or the 'within-cluster sum of square errors' (WSS), represents the dispersion of the attribute values of individual objects within the clusters they belong to (Zhang et al., 2016). The procedure starts with an initial partitioning (usually created randomly) and then continues with modifying it iteratively until the best possible division into groups is found, i.e. the smallest WSS possible is achieved. In other words, at each step, increasingly more homogeneous groups are created. As a result of the algorithm implementation, we obtain a division of the set into a specified number of clusters and the value of WSS, which characterises the final

[7] A draft manuscript of this paper first circulated for comments in 1957 at Bell Telephone Laboratories, but was submitted for publication only in 1981.

level of the clustering quality (the lower the WSS, the better the clustering). Intuitively, it may be expected that the larger the number of clusters, the lower the WSS value. However, from the practical point of view, large numbers of clusters (groups) may not always be useful for researchers. They may cope with cluster abundance by assuming a maximum number of groups that would suit their research purposes. The challenge of selecting the optimal number of clusters will be addressed later in this section. Calculations for the purpose of the analysis were made with the kmeans ('stats' package) function of the R software (R Core Team, n.d.), implementing the *k*-means algorithm by default in the Hartigan and Wong version (Hartigan & Wong, 1979).

The *k*-means has been recently applied in various research topics in regional studies (Table 2), serving to create regionalisations as well as classifications and typologies.

Table 2. Examples of the application of *k*-means clustering in regional studies

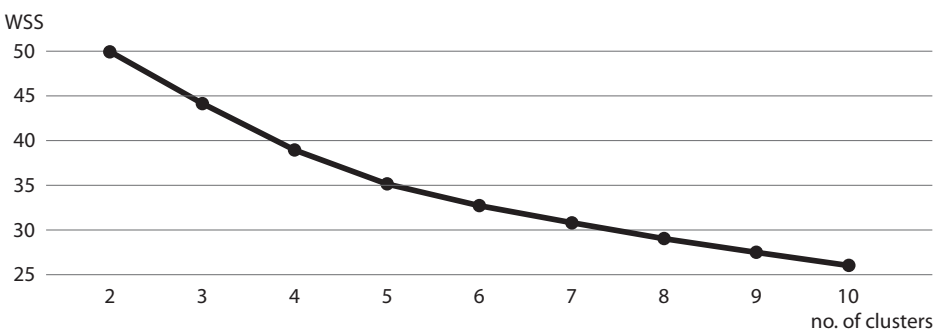
Paper	Purpose of the grouping
Crone (2005)	regionalisation of the US states (48 contiguous states) in terms of business cycles – comparing own clustering to the one performed by the Bureau of Economic Analysis (BEA)
Brauksa (2013)	exploration of socio-economic differentiation of Latvian municipalities following an administrative territorial reform
Šebová (2013), as cited by Mikuš (2016)	regionalisation of Slovakia in terms of socio-economic marginality
Li et al. (2016)	*k*-means used to support a geographically weighted regression (GWR) analysis of community resilience to seismic hazards in Southwest China, in particular to divide the study area into subregions based on characteristics defined by regression coefficients
Malinowski (2016)	classification of Polish regions in what concerns a relationship between the human potential and the economic effectiveness of enterprises – partitioning around medoids (PAM, *k*-medoids) method instead of *k*-means (a similar non-hierarchical clustering algorithm)
Novotná et al. (2016)	socioeconomic spatial typology of the Pilsen region, the Czech Republic
Stukalo & Simakhova (2018)	investigation of social economy patterns for 40 selected countries
Bole et al. (2019)	typology of small industrial towns in Slovenia
Dawidowicz (2020)	assessment of the financial situation of Polish gminas
Bayisa et al. (2020)	modelling and forecasting ambulance calls in Northern Sweden: *k*-means clustering used for the estimation of one of the parameters of the model
Kong et al. (2021)	exploring energy consumption patterns within the Qinghai Province, China

Source: authors' work based on a literature review.

Not having any initial assumption as to the desired number of clusters, we tested grouping variants from 2 to 10 clusters, and then evaluated the outcome based on the WSS value. Then we decided on how many clusters to consider using the elbow

method (Everitt et al., 2011; Peeples, 2011; Thorndike, 1953; Tibshirani et al., 2001; Zhang et al., 2016). This approach consists in the visual analysis of a graph showing the dependence between the WSS value and the number of clusters: we examine the course of the WSS (k) function (the WSS value for the division of the studied group of objects into k clusters) and search for such k (number of clusters) for which the graph clearly breaks down. The very idea of this method dates back to the 1950s (Thorndike, 1953) and is based on theoretical considerations related to the 'Gap statistic' (Tibshirani et al., 2001). As a result of its application, taking into account the graph obtained on the basis of our calculations (Figure 3), we came to the conclusion that the optimal division of our set involves five clusters.

Figure 3. The WSS of the k-means cluster analysis performed on a set of 192 gminas of the Polish Carpathians



Source: authors' work based on an analysis performed in R software involving data from the Local Data Bank of Statistics Poland.

While the elbow method can be seen as a bit arbitrary approach, we compared its outcome with the Calinski-Harabasz index (also called the pseudo-F statistic), which is a ratio reflecting within-group similarity and between-group difference (Caliński & Harabasz, 1974; Milligan & Cooper, 1985), and obtained the same results.

3. Results: the final selection of gminas for case studies

In the process of the final selection of the gminas for case studies, we adopted further criteria to narrow the potential set of the entities. Firstly, we decided to consider only those gminas which have national parks (or at least their buffer zone) and Natura 2000 site(s) other than parks[8] on their territory in order to identify the similarities and differences between these types of protected areas. Subsequently, we excluded

[8] In Poland, all national parks are by default included in the Natura 2000 network, so only analysing the Natura 2000 sites besides parks made sense in this research.

smaller towns and all gminas of the Tatra Mountains. The latter would be a too specific case as the area forms the top tourist destination in Poland, while we were interested in discovering any subtle interrelationships between nature protection and development; additionally, we aimed to focus on more peripheral areas. Next, we looked at how the 23 gminas which met the criteria above were distributed among the five groups determined on the basis of *k*-means clustering.

The default approach involved selecting one case per group. For this purpose, we examined the positions of the gminas within their groups, understood as the Euclidean distance from the groups' centres (Figure 4). We preferred to choose those cases which were close to their clusters' centres when one case was to be selected or, by contrast, to be not too close to each other (but also not too far from the cluster centre) when two cases in one group were considered. These analyses (Figure 4B) also led us to refrain from selecting a representative of cluster 4, which contained only two gminas that met our additional criteria described above – Czorsztyn was here a big outlier, while Krościenko nad Dunajcem appeared to us a too similar case to Szczawnica in terms of its location (Pieniny Mts), and by contrast, not having any additional features (Table 3) which were interesting from the perspective of the purpose of the study. The principle component analysis (PCA) demonstrated that clusters 2 and 4 became clearly separated only after the PC3 was used (the PC1 to PC2 diagram in Figure 4A shows that the two groups were still quite mixed), which suggests that cluster 4 (which incidentally formed the smallest group) can be to some extent treated as a subgroup of cluster 2. The latter circumstance additionally justified Szczawnica being chosen instead of any of the other two Pieniny gminas which fell into cluster 4.

The choice described above was further confirmed by the results of an analysis of the clusters' characteristics, involving both an investigation of the distribution of the variables considered for the *k*-means procedure (Figure 5) and a general, more general knowledge about the studied area, including qualitative and difficult to measure information.

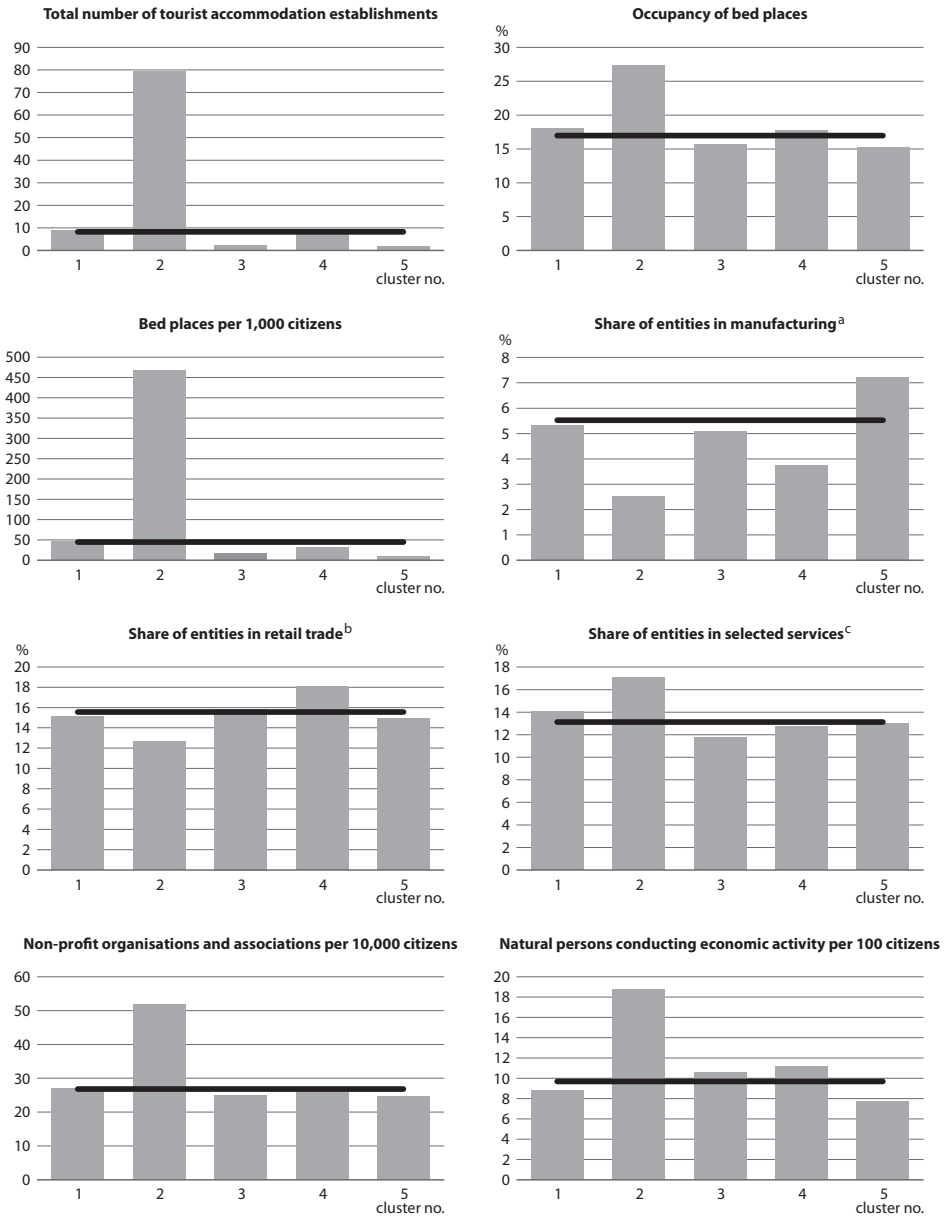
The latter included such aspects as the social reception of the protected areas (examples of both conflicts and good cooperation practices, resulting from the analysis of public consultation documents and from the local press[9]). We aimed at creating a well-balanced representation of different Carpathian physical and cultural landscapes and strived to include at least one gmina specialising in spa treatment (Table 3) in order to verify whether this kind of 'gmina profile' goes hand in hand with nature protection, as simple logic would suggest.

[9] For more details on the sources used at this stage of the project, please see Warchalska-Troll (2018, pp. 51–52).

Figure 4. Preselected gminas of the Polish Carpathians distributed among five clusters

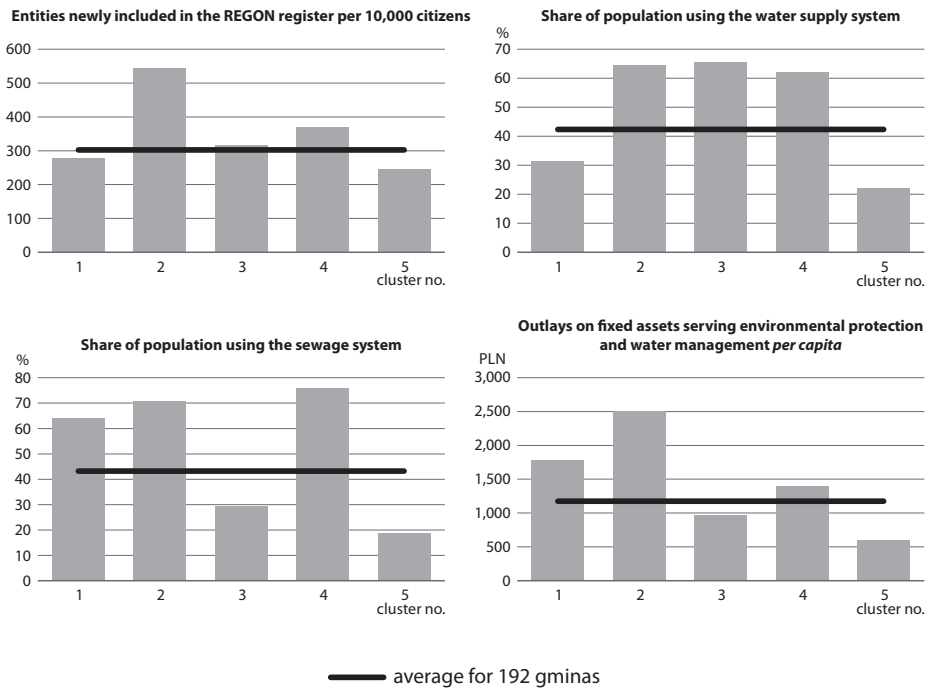
Note. The eventually selected gminas are shown in boxes.
Source: authors' work based on the analysis performed in R software using data from the Local Data Bank of Statistics Poland.

Figure 5. Average values for the variables considered in *k*-means clustering of 192 gminas of the Polish Carpathians for the five selected clusters, compared to the average values for the whole set



a Share of entities in section C, divisions 10, 11, 15 and 16 of NACE Rev. 1.2 / PKD 2007 in the total number of entities in the REGON register. b Share of entities in section G, division 47 of NACE Rev. 1.2 / PKD 2007 (except for the sales of motor vehicles and motorcycles) in the total number of entities in the REGON register. c Share of entities in sections/divisions I56, M72, 74, P85, R90, 91, 93 and S94 of NACE Rev. 1.2 / PKD 2007 in the total number of entities in the REGON register.

Figure 5. Average values for the variables considered in *k*-means clustering of 192 gminas of the Polish Carpathians for the five selected clusters, compared to the average values for the whole set (cont.)



Note. The data for the analysis were obtained in early 2014 and show the situation as of 2012, except for entities newly added to the REGON register per 10,000 citizens (which include the total for the whole available period, i.e. 2009–2012), and outlays on fixed assets serving environmental protection and water management *per capita* (the total for the whole available period, i.e. 1999–2008). For more information on the data used, please see Table 1.
Source: authors' work based on data from the Local Data Bank of Statistics Poland.

Table 3. A brief description of the main characteristics of the five clusters obtained through *k*-means clustering

Cluster no.	Cluster characteristics	Distribution of the 23 gminas selected at earlier stages	Long-lasting and significant conflicts over nature protection	Grounded and wide cooperation between the local people and nature protection stakeholders
1	Urban-rural or rural gminas retaining an agricultural component and traditional activities/lifestyles/folklore	Jabłonka	no	no
		Kamienica	no	no
		Krempna	yes	no
		Lipinki	no	no
		Lipnica Wielka	no	yes
		Łapsze Niżne	no	no
		Ochotnica Dolna	yes	yes

Table 3. A brief description of the main characteristics of the five clusters obtained through *k*-means clustering (cont.)

Cluster no.	Cluster characteristics	Distribution of the 23 gminas selected at earlier stages	Long-lasting and significant conflicts over nature protection	Grounded and wide cooperation between the local people and nature protection stakeholders
2	Largest tourist centres	**Cisna**	yes	no
		Lutowiska	no	yes
		Solina	no	no
		Szczawnica[a]	yes	no
3	Local trade and services centres with no single specialisation	Dukla	no	no
		Nowy Targ	no	no
		Sękowa[a]	yes	yes
4	Urban gminas and secondary tourist centres	Czorsztyn	no	no
		Krościenko nad Dunajcem	no	no
5	Peripheral, rural gminas with an important production (mostly forestry and logging) component, looking for their opportunity in winter sports with moderate success	Czarna	no	no
		Dębowiec	no	no
		Mszana Dolna	no	no
		Niedźwiedź	no	no
		Nowy Żmigród	no	no
		Osiek Jasielski	no	no
		Zawoja	yes	no

a Spa resorts.

Note. The table illustrates the situation as of 2014.

Source: authors' work based on an analysis of the official documents relating to public consultations (concerning plans issued by nature protection institutions) and the local press.

We not only investigated how the 23 preselected gminas were distributed within the clusters, but also their within-group positions (Figure 4). Then the spatial autocorrelation was examined to reveal which gminas significantly differ from their neighbourhood, and as such may provide some additional input. For this purpose the Anselin Local Morans I tool available in the ArcGIS software was used. We counted how many times each gmina stood out from the entire considered set at this stage, taking into account some basic features (population density, tourism infrastructure, transport accessibility, environmental infrastructure, economic and civic/social activity of the population, poverty). As a result, the Cisna gmina joined the already selected Szczawnica gmina in cluster 2. Cisna was the only gmina which appeared different from its neighbours in terms of five inquiries, five times showing a 'high-high' autocorrelation. To justify two cases appearing also in cluster 1, given the general profile of this cluster (urban-rural or rural gminas retaining an agricultural component and traditional activities/lifestyles/folklore (Table 3), the

authors chose Ochotnica Dolna and Lipinki, which represent two different regions (Gorce Mts and Beskid Niski Mts, respectively), two different physical features (Ochotnica Dolna is located in a mountain valley with steep slopes and harsh climate while Lipinki are situated in the foothills, with much more favourable conditions for agriculture), and, as a result, two different agricultural specialisations (pastoralism with mostly sheep herding and crops growing and cattle farms, respectively). Moreover, Ochotnica Dolna has a history of intensive and yet mixed experiences in cooperating with nature protection institutions. On the one hand, the gmina had a relatively good relationship with a nearby national park, based on common projects relating to the promotion of local cultural heritage and traditional summer farming revival, but on the other hand, numerous conflicts occurred in the process of implementing the Natura 2000 network on the gmina's territory. In turn, Lipinki had the lowest in the whole set values of variables relating to the development of tourism and as such effectively balanced the tourism-oriented gminas of Szczawnica, Cisna, Zawoja and (to a certain extent) Sękowa in our final case study set.

Among the numerous clustering techniques (Everitt et al., 2011), the *k*-means algorithm can be perceived as relatively time-effective and intuitive and its results can be easily interpreted; therefore, it may be a good choice when a quick outcome is anticipated. In this research, the overall scheme (summarised in Figure 2), whose central element was *k*-means partitioning, not only confirmed the authors' initial intuition of which gminas could be interesting to take a closer look at (e.g. Szczawnica, Zawoja), but also suggested the less obvious choices of Sękowa or Lipinki, and indicated Cisna as the most specific case in the generally 'standing out' region of the Bieszczady Mts. Moreover, as noted by Zhang et al. (2016), in contrast to many other methods (e.g. hierarchical ones), this one is less vulnerable to noisy data and was found to produce more stable results (i.e. a slight change in the scope of the data does not significantly affect the clustering result). Its certain inconvenience – arbitrarily assigning the number of clusters as its parameter – is in most cases easy to overcome by repeating the calculations for other values (it is usually quite clear for the researcher to what extent the number of clusters may vary). In these kinds of instances, the best option can be determined on the basis of one of the methods of evaluating the quality of clustering (Everitt et al., 2011; Kodinariya & Makwana, 2013; Migdał-Najman, 2011; Milligan & Cooper, 1985), like in our case the elbow method (Thorndike, 1953; Tibshirani et al., 2001) and the pseudo-*F* statistic, also called the Calinski-Harabasz index (Caliński & Harabasz, 1974; Everitt et al., 2011; Larose & Larose, 2014; Milligan & Cooper, 1985).

When reviewing studies similar to ours, we have observed that the number of clusters was sometimes based on an expert decision preceded by a qualitative or

quantitative analysis of the study area (e.g. Novotná et al., 2016), apart from cases when the amount of clusters was defined in advance due to a research hypothesis or prerequisites (e.g. Brauksa, 2013; Crone, 2005; Malinowski, 2016). In contrast, some authors propose to derive the optimal number of clusters from hierarchical clustering (e.g. Bole et al., 2019; Šebová, 2013 as cited by Mikuš et al., 2016; Stukalo & Simakhova, 2018) or from the Akaike information criterion (AIC) method (Li et al., 2016). Not undermining any of these approaches, similarly to e.g. Dawidowicz (2020), Kong et al. (2021), Nicholson et al. (2019)[10] and Zhang et al. (2016),[11] we opted for the elbow method which is simple to use and intuitive. Its possible drawback is that the curve which is supposed to indicate the optimal number of clusters may sometimes appear too 'smooth' (not having a distinct 'elbow'). This can happen when a large number of observations is considered and a relatively small number of clusters may be accepted. A possible solution to this problem could be replacing the WSS parameter by the difference in minWSS (Zhang et al., 2016). Another option is to verify the results obtained through the elbow method with an alternative technique, as we did by means of the pseudo-*F* statistic.

While using the *k*-means algorithm one should bear in mind that it may produce slightly different outcomes depending on which initial centroids are selected, as explained by Everitt et al. (2011) and noticed by many researchers (Crone, 2005; Novotná et al., 2016; Zhang et al., 2016). So, in case of doubts and if a particular implementation allows it, calculations may be repeated with other initial centroids (which is possible using the R software as we did). Gao & Kupfer (2018) also emphasise the fact that the *k*-means algorithm is 'non-spatial', as it does not seek to define by default the spatially coherent regions. This was not a problem in our case, as we only needed to classify gminas, not find clusters in the spatial sense. However, in some instances the latter may be achieved by adding extra attributes which numerically define the 'levels of neighbourhood' between the objects of research (Crone, 2005). Thanks to this possibility, the *k*-means method is likely to become more broadly appreciated by geographers. Finally, we would like to mention that apart from the classic GIS software, algorithms dedicated to solving tasks relating to data analysis in a spatial context are also available in the R software we used (e.g. the skater function of the spdep package as shown by Nicholson et al., 2019). In turn, the popular GIS software, ESRI ArcGIS, applies the *k*-means algorithm for its Grouping Analysis tool with a 'no spatial constraint' parameter (ESRI, n.d.).

[10] Nicholson et al. (2019) use the skater function in the spdep R package (instead of the *k*-means), which also performs spatial clustering and the value of the *k* parameter is to be defined by the researcher.
[11] Zhang et al. (2016) use a modified variant of the elbow method.

4. Conclusions

To sum up, in the procedure we followed, we combined a statistical analysis with qualitative criteria to narrow and finalise a selection of gminas for in-depth studies, as we believe that statistics should verify and support, not depreciate or dominate the 'spatial intuition' of geographers. The fact that we operated on qualitative and quantitative data of a diverse quality, and broadly used official statistics data, proves that our observations are not detached from the real challenges faced by scholars in this field. Our calculations were made using the open source R software, which does not have any specific requirements concerning the hardware, so our scheme may be copied by everyone. We did not want our research to be consistent with the belief that although statistics reveal a lot, at the same time they tend to conceal what is vital, therefore we considered a wide variety of non-measurable factors at the final stage of our research framework.

In conclusion, we have noticed a considerable potential of *k*-means clustering for case study selection in geographical research, which so far seems to have been only superficially explored in this application. In contrast to what is sometimes practised, we would not recommend this technique for regionalisation and other analyses aiming to find spatial clusters (or at least not this technique alone, as by definition it is non-spatial) but rather for classifications and typologies. When deciding which area to study, this method is likely to remain objective and not limited by measurability when supplemented by qualitative information concerning the local and regional spatial contexts. A practicable improvement to the analysis scheme described in this paper may concern the use of principal components instead of raw data for less correlated and more refined variables, similarly for instance to what Bole et al. (2019) described. Future studies relating to the topic may also include further research as to what extent and under what conditions clusters obtained from different partitioning methods converge.

References

Babbie, E. (2007). *Badania społeczne w praktyce*. Wydawnictwo Naukowe PWN.

Bayisa, F. L., Ådahl, M., Rydén, P, & Cronie, O. (2020). Large-scale modelling and forecasting of ambulance calls in northern Sweden using spatio-temporal log-Gaussian Cox processes. *Spatial Statistics*, *39*, 1–22. https://doi.org/10.1016/j.spasta.2020.100471.

Bole, D., Kozina, J, & Tiran, J. (2019). The variety of industrial towns in Slovenia: a typology of their economic performance. *Bulletin of Geography. Socio-economic Series*, *46*(46), 71–83. http://doi.org/10.2478/bog-2019-0035.

Brauksa, I. (2013). Use of Cluster Analysis in Exploring Economic Indicator Differences among Regions: The Case of Latvia. *Journal of Economics, Business and Management*, *1*(1), 42–45. http://doi.org/10.7763/JOEBM.2013.V1.10.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27.

Crone, T. M. (2005). An alternative definition of economic regions in the United States based on similarities in state business cycles. *The Review of Economics and Statistics*, *87*(4), 617–626. https://doi.org/10.1162/003465305775098224.

Dawidowicz, D. (2020). Ocena sytuacji finansowej gmin z wykorzystaniem metody *k*-średnich. *Wiadomości Statystyczne. The Polish Statistician*, *65*(7), 26–46. https://doi.org/10.5604/01.3001.0014.3284.

ESRI. (n.d.). *Grouping Analysis (Spatial Statistics) 8.* Retrieved June 24, 2021, from https://pro.arcgis.com/en/pro-app/2.8/tool-reference/spatial-statistics/grouping-analysis.htm.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis* (5th edition). John Wiley & Sons. https://doi.org/10.1002/9780470977811.

Gao, P., & Kupfer, J. A. (2018). Capitalizing on a wealth of spatial information: Improving biogeographic regionalization through the use of spatial clustering. *Applied Geography*, *99*, 98–108. https://doi.org/10.1016/j.apgeog.2018.08.002.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A *K*-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108. https://doi.org/10.2307/2346830.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, *1*(6), 90–95.

Kong, W., Wang, Y., Dai, H., Zhao, L., & Wang, C. (2021). Analysis of energy consumption structure based on K-means clustering algorithm. *E3S Web of Conferences*, *267*, 1–5. https://doi.org/10.1051/e3sconf/202126701054.

Kraszewska, B. (2016). Wykorzystanie analizy skupień w ocenie zróżnicowania zagrożenia ubóstwem w podregionach Polski. *Wiadomości Statystyczne. The Polish Statistician*, *61*(5), 17–36. https://doi.org/10.5604/01.3001.0014.0993.

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining* (2nd edition). John Wiley & Sons. https://doi.org/10.1002/9781118874059.

Li, X., Wang, L., & Liu, S. (2016). Geographical Analysis of Community Resilience to Seismic Hazard in Southwest China. *International Journal of Disaster Risk Science*, *7*(3), 257–276. https://doi.org/10.1007/s13753-016-0091-8.

Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297). University of California Press. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Some-methods-for-classification-and-analysis-of-multivariate-observations/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992.

Malinowski, M. (2016). Potencjał ludzki a efektywność ekonomiczna przedsiębiorstw – wykorzystanie metod taksonomicznych w ujęciu regionalnym. *Studia Regionalne i Lokalne*, (2), 87–109. https://doi.org/10.7366/1509499526405.

Migdał-Najman, K. (2011). Ocena jakości wyników grupowania – przegląd bibliografii. *Przegląd Statystyczny*, *58*(3–4), 281–299.

Mikuš, R., Máliková, L., & Lauko, V. (2016). An introductory study of perceptual marginality in Slovakia. *Bulletin of Geography. Socio-economic Series*, (34), 47–62. http://dx.doi.org/10.1515/bog-2016-0034.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179. https://doi.org/10.1007/BF02294245.

Nicholson, D., Vanli, O. A., Jung, S., & Ozguven, E. E. (2019). A spatial regression and clustering method for developing place-specific social vulnerability indices using census and social media data. *International Journal of Disaster Risk Reduction*, *38*, 101–224 https://doi.org/10.1016/j.ijdrr.2019.101224.

Novotná, M., Šlehoferová, M., & Matušková, A. (2016). Evaluation of spatial differentiation in the Pilsen region from a socio-economic perspective. *Bulletin of Geography. Socio-economic Series*, (34), 73–90. https://doi.org/10.1515/bog-2016-0036.

Peeples, M. A. (2011). *R Script for K-Means Cluster Analysis*. Retrieved May 27, 2021, from http://www.mattpeeples.net/kmeans.html.

R Core Team. (n.d.). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Retrieved August 30, 2020, from https://www.R-project.org/.

Steinhaus, H. (1957). Sur la division des corps matériels en parties. *Bulletin L'Académie Polonaise des Sciences*, *4*(12), 801–804. http://www.laurent-duval.eu/Documents/Steinhaus_H_1956_j-bull-acad-polon-sci_division_cmp-k-means.pdf.

Stukalo, N., & Simakhova, A. (2018). Global parameters of social economy clustering. *Problems and Perspectives in Management*, *16*(1), 36-47. https://doi.org/10.21511/ppm.16(1).2018.04.

Taylor, L. (2016). Case Study Methodology. In N. Clifford, M. Cope, T. Gillespie & S. French (Eds.), *Key Methods in Geography* (3rd edition; pp. 581–595). SAGE Publications.

Thorndike, R. L. (1953). Who belongs in the family?. *Psychometrika*, *18*(4), 267–276. https://doi.org/10.1007/BF02289263.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293.

Warchalska-Troll, A. (2018). Natura 2000 sites in the Polish Carpathians vs local development: inevitable conflict?. *eco.mont Journal of Mountain Protected Areas Research and Management, 10*(2), 50–58. https://doi.org/10.1553/eco.mont-10-2s50.

Warchalska-Troll, A. (2019). Do Economic Opportunities Offered by National Parks Affect Social Perceptions of Parks? A Study from the Polish Carpathians. *Mountain Research and Development, 39*(1), 37–46. https://doi.org/10.1659/MRD-JOURNAL-D-18-00055.1.

Zhang, Y., Moges, S., & Block, P. (2016). Optimal Cluster Analysis for Objective Regionalization of Seasonal Precipitation in Regions of High Spatial–Temporal Variability: Application to Western Ethiopia. *Journal of Climate*, *29*(10), 3697–3717. https://doi.org/10.1175/JCLI-D-15-0582.1.

Wiadomości Statystyczne. The Polish Statistician, 2022, vol. 67, 2, 21–35 DOI: 10.5604/01.3001.0015.7718
Edukacja statystyczna / Statistical education
Zgłoszony/submitted: 6.09.2021, poprawiony/revised: 21.01.2022, zaakceptowany/accepted: 13.02.2022

O różnych kryteriach konstrukcji przedziałów ufności na przykładzie rozkładu normalnego

Stanisław Jaworski[a]

Streszczenie. W artykule opisano różne kryteria konstrukcji przedziałów ufności oparte na kontroli błędu przeszacowania i niedoszacowania parametru, jak również na zadanej z góry precyzji oszacowania tego parametru. Odpowiednie konstrukcje zilustrowano przedziałami ufności dla parametrów rozkładu normalnego. Zamieszczone w pracy przykłady pokazują wpływ zadanych kryteriów na postać przedziałów ufności, a wykonane obliczenia przedstawiają zależność szerokości danego przedziału od rozmiaru próby i odwrotnie – rozmiaru próby od zadanej szerokości tego przedziału. Zagadnienie doboru wielkości próby przy zadanej z góry precyzji oszacowania parametru zaprezentowano z uwzględnieniem rozwiązań, które nie zależą od szacunkowych wartości nieznanego parametru. Celem artykułu jest pokazanie, w jaki sposób można rozszerzyć temat konstrukcji przedziałów ufności w edukacji statystycznej.

Słowa kluczowe: dobór wielkości próby, poziom ufności, precyzja oszacowania, przedziały ufności

JEL: C13

On various criteria for the construction of confidence intervals on the example of normal distribution

Abstract. The paper describes various criteria for the construction of confidence intervals based on the control of the error of overestimation or underestimation of a parameter and on the pre-determined precision of the parameter estimation. The respective constructions are illustrated by confidence intervals for the parameters of the normal distribution. Examples presented in the paper show the influence of some selected criteria on the formula of the appropriate confidence interval. The performed calculations show the dependence of the width of a given interval on the size of the sample, and vice versa – the size of the sample on the pre-determined width of this interval. The problem of selecting the sample size with a pre-determined precision of the estimation was presented taking into account solutions that do not depend on the estimated values of the unknown parameter. The aim of the study is to show how the problem of confidence intervals construction can be presented more comprehensively in statistical education.

Keywords: sample size selection, confidence level, precision of estimation, confidence intervals

[a] Szkoła Główna Gospodarstwa Wiejskiego w Warszawie, Wydział Zastosowań Informatyki i Matematyki, Katedra Ekonometrii i Statystyki, Polska / Warsaw University of Life Sciences, Faculty of Applied Informatics and Mathematics, Department of Econometrics and Statistics, Poland. ORCID: https://orcid.org/0000-0002-6169-2886. E-mail: stanislaw_jaworski@sggw.edu.pl.

1. Wprowadzenie

Przedziały ufności, wprowadzone do wnioskowania statystycznego przez Jerzego Spławę-Neymana, należą do najważniejszych narzędzi statystycznych. Zasadnicza praca o przedziałach ufności pochodzi z 1937 r. (Neyman, 1937), aczkolwiek to wcześniejsza publikacja – artykuł zamieszczony w „Journal of the Royal Statistical Society" (Neyman, 1934) – została uznana za jedno z największych dwudziestowiecznych osiągnięć w dziedzinie metodologii statystyki (Kotz i Johnson, 1992). Przełomowy artykuł z 1934 r. powstał na bazie monografii z 1933 r. opracowanej w ramach współpracy Neymana z Instytutem Spraw Społecznych (Krzyśko i in., 2018).

Neyman (1934) zaproponował dwa kryteria konstrukcji przedziału ufności:

1. przedział musi dać się wyznaczyć;
2. przedział powinien być jak najwęższy.

Pierwsze kryterium wyraża pogląd Neymana, że statystyka powinna mieć wymiar praktyczny. Drugie kryterium, wynikające z postulatu, że błąd estymacji powinien być jak najmniejszy, można modyfikować poprzez uwzględnienie z jednej strony ryzyka przeszacowania i niedoszacowania parametru, z drugiej zaś – doboru wielkości próby gwarantującej zadaną precyzję oszacowania. Te modyfikacje prowadzą do kolejnych interesujących kryteriów konstrukcji przedziałów ufności, które są przedmiotem niniejszego artykułu i zostaną omówione, a następnie zilustrowane na przykładzie modelu gaussowskiego. Zamieszczone obliczenia uwzględniają praktyczne możliwości ich realizacji i pokazują wpływ, jaki na konstrukcję przedziału ufności ma kontrola jego szerokości oraz błędu polegającego na przeszacowaniu lub niedoszacowaniu parametru. Wynika z tego, że praktyczne zastosowanie konkretnego przedziału ufności powinno być zawsze poprzedzone pytaniem, czy przedział ten jest odpowiedni dla badanego zagadnienia oraz czy rozmiar próby, którą dysponujemy, jest wystarczający do uzyskania pożądanej precyzji oszacowań. Dla praktyków te pytania są bardzo ważne, dlatego wydaje się, że uwzględnienie omówionych konstrukcji w edukacji statystycznej jest jak najbardziej zasadne.

2. Przedział ufności

Podstawą konstrukcji przedziału ufności jest n-wymiarowy wektor losowy (wektor obserwacji) $\boldsymbol{X}$ o rozkładzie prawdopodobieństwa P_θ. Rozkład ten zależy od nieznanego parametru θ. Wiadomo jedynie, że parametr ten należy do pewnego znanego zbioru Θ, nazywanego *przestrzenią parametrów*. Na podstawie realizacji wektora losowego $\boldsymbol{X}$ można oszacować ten parametr za pomocą przedziału ufności. Klasyczna definicja *przedziału ufności* jest następująca:

Definicja. Przedział $(\theta_L(\boldsymbol{X}), \theta_R(\boldsymbol{X}))$, taki że

$$P_\theta\{\theta \in \big(\theta_L(\boldsymbol{X}), \theta_R(\boldsymbol{X})\big)\} \geq \delta, \forall\theta \in \Theta, \tag{1}$$

nazywamy przedziałem ufności dla parametru θ, na poziomie ufności $\delta \in (0,1)$.

W praktyce przyjmuje się, że δ jest ustaloną wartością, najczęściej równą 0,95 lub 0,99. Podana definicja dotyczy parametru jednowymiarowego. Gdyby parametr θ był wielowymiarowy, podana definicja nie zmieniłaby się, poza tym że słowo *przedział* zostałoby zastąpione słowem *zbiór* lub *obszar*. W podanej definicji prawdopodobieństwo $P_\theta\big\{\theta \in \big(\theta_L(\boldsymbol{X}), \theta_R(\boldsymbol{X})\big)\big\}$ jest rozumiane następująco:

$$P_\theta\left\{\boldsymbol{X} \in \left\{\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{n}}\colon \theta \in \big(\theta_L(\boldsymbol{x}), \theta_R(\boldsymbol{x})\big)\right\}\right\}. \tag{2}$$

Oznacza to, że w przeciwieństwie do wektora $\boldsymbol{X}$ parametr θ nie jest losowy. Dlatego mówimy o prawdopodobieństwie pokrycia parametru θ przez przedział ufności, a nie o prawdopodobieństwie, że parametr θ należy do tego przedziału. Chociaż można uznać, że obie interpretacje są logicznie równoważne, to druga sugeruje, że θ jest zmienną losową, co nie jest prawdą. Przypadek modelu statystycznego, w którym θ jest zmienną losową, prowadzi do pojęcia *bayesowskiego przedziału wiarygodności*.

3. Kryteria konstrukcji przedziału ufności

Wnioskując na podstawie przedziału ufności, gdy wektor losowy $\boldsymbol{X}$ zrealizuje się jako $\boldsymbol{x} \in \mathbb{R}^n$, możemy mieć do czynienia z jedną z trzech niewidocznych dla nas sytuacji:
1. $\theta \in \big(\theta_L(\boldsymbol{x}), \theta_R(\boldsymbol{x})\big)$;
2. $\theta \leq \theta_L(\boldsymbol{x})$;
3. $\theta \geq \theta_R(\boldsymbol{x})$.

W każdej z nich wyprowadzamy ten sam wniosek, że prawdziwa i nieznana wartość parametru θ została pokryta przez przedział $\big(\theta_L(\boldsymbol{x}), \theta_R(\boldsymbol{x})\big)$. Dlatego jeżeli zajdzie pierwsza sytuacja, wyprowadzony wniosek jest poprawny. W drugiej mówimy o błędzie przeszacowania, a w trzeciej – o błędzie niedoszacowania. Przy poziomie ufności δ ryzyko popełnienia błędu przeszacowania lub niedoszacowania wynosi co najwyżej $1-\delta$. Jeżeli oznaczymy ryzyko przeszacowania jako α_p, a ryzyko niedoszacowania jako α_n, to zgodnie z ideą konstrukcji przedziału ufności mamy:

$$\alpha_p + \alpha_n \leq 1 - \delta. \tag{3}$$

Formalnie $\alpha_p = P_\theta\{\theta \leq \theta_L(\boldsymbol{X})\}$ i $\alpha_n = P_\theta\{\theta \geq \theta_R(\boldsymbol{X})\}$.

Najpopularniejszym kryterium konstrukcji przedziału ufności jest symetryczny podział ryzyka błędu. Zgodnie z tym kryterium funkcje $\theta_L(\boldsymbol{x})$ oraz $\theta_R(\boldsymbol{x})$ dobiera się tak, aby zachodziła równość $P_\theta\{\theta \leq \theta_L(\boldsymbol{X})\} = P_\theta\{\theta \geq \theta_R(\boldsymbol{X})\}$. W tym przypadku ryzyko przeszacowania i ryzyko niedoszacowania są takie same: $\alpha_p = \alpha_n$. W przypadku gdy wektor losowy $\boldsymbol{X}$ ma rozkład typu ciągłego, można przyjąć $\alpha_p = \alpha_n = (1-\delta)/2$.

Zdarzają się sytuacje, w których z powodów merytorycznych bardziej wskazany jest niesymetryczny podział ryzyka błędu. Na przykład nie chcąc popełnić błędu przeszacowania, należy przyjąć $\alpha_p = 0$. Konsekwencją takiego założenia jest przyjęcie $\theta_L \equiv \inf\Theta$. Jeżeli parametr θ jest średnią w rozkładzie normalnym, wówczas $\theta_L \equiv -\infty$, a jeżeli oznacza prawdopodobieństwo sukcesu w rozkładzie dwumianowym, to $\theta_L \equiv 0$. Innymi słowy, nie ma niebezpieczeństwa przeszacowania, ponieważ lewy koniec przedziału zawsze znajdzie się poniżej nieznanej wartości parametru, bez względu na obserwacje otrzymane w doświadczeniu. Uzyskujemy przedział, w którym wyznaczany jest tylko jego prawy koniec. Taki przedział ufności nazywany jest *jednostronnym przedziałem ufności* (dokładniej: *prawostronnym*). Przedział lewostronny uzyskuje się dla $\alpha_n = 0$. Wtedy $\theta_R \equiv \sup\Theta$, co oznacza brak możliwości niedoszacowania parametru.

Bardzo istotnym elementem we wnioskowaniu opartym na przedziałach ufności jest, wskazana przez Neymana (1934), szerokość przedziału ufności. Dwa wyżej wymienione kryteria kładły nacisk jedynie na ryzyko błędów, co nie zapewniało uzyskiwania najwęższych przedziałów. Minimalizacja szerokości przedziału ufności polega na takim doborze funkcji $\theta_L(\boldsymbol{x})$ i $\theta_R(\boldsymbol{x})$, aby dla wszystkich możliwych obserwacji $\boldsymbol{x}$ szerokość przedziału, tzn. różnica $\theta_R(\boldsymbol{x}) - \theta_L(\boldsymbol{x})$, była najmniejsza. Gdy nie jest to możliwe, można przyjąć nieco słabsze kryterium minimalizacji przeciętnej szerokości przedziału ufności, czyli taki dobór funkcji $\theta_L(\boldsymbol{x})$ i $\theta_R(\boldsymbol{x})$, dla których wielkość

$$E_\theta(\theta_R(\boldsymbol{X}) - \theta_L(\boldsymbol{X})) \tag{4}$$

jest najmniejsza dla każdego $\theta \in \Theta$.

Na podstawie szerokości przedziału ufności można przyjąć kryterium, którego spełnienie zależy nie tyle od końców przedziału ufności, ile od rozmiaru próby. Nawet najwęższy przedział ufności może w praktyce nie być dostatecznie wąski. Dlatego też rozważa się dwa podejścia polegające na konstrukcji przedziałów ufności o zadanej szerokości, w szczególności takiej, że:

1. średnia szerokość przedziału ufności ma nie przekraczać zadanej wartości $d > 0$, tzn.

$$E_\theta\big(\theta_R(\boldsymbol{X}) - \theta_L(\boldsymbol{X})\big) \leq d, \forall\theta \in \Theta; \tag{5}$$

2. z dużym prawdopodobieństwem szerokość przedziału ufności ma nie przekraczać zadanej wartości $d > 0$, czyli

$$P_\theta\{\theta_R(\boldsymbol{X}) - \theta_L(\boldsymbol{X}) \leq d\} \geq \gamma, \forall\theta \in \Theta, \tag{6}$$

gdzie γ jest daną wartością bliską 1.

4. Konstrukcja przedziału ufności

Zasadnicza konstrukcja przedziału ufności polega na znalezieniu takiego zbioru A_θ, dla którego zachodzi $P_\theta\{\boldsymbol{X} \in A_\theta\} \geq \delta$, dla każdej wartości parametru $\theta \in \Theta$. Następnie dla dowolnej realizacji $\boldsymbol{x}$ wektora losowego $\boldsymbol{X}$ definiuje się następujący zbiór:

$$C(\boldsymbol{x}) = \{\theta \in \Theta: \boldsymbol{x} \in A_\theta\}. \tag{7}$$

Zbiór losowy $C(\boldsymbol{X})$ jest zbiorem ufności na poziomie ufności δ. W praktyce dla jednowymiarowego parametru θ przyjmuje on zazwyczaj postać przedziału.

Przy konstrukcji przedziału ufności przydaje się funkcja centralna $t(\boldsymbol{x}, \theta)$ (Krzyśko, 1996). Zgodnie z definicją tej funkcji $t(\boldsymbol{X}, \theta)$ jest zmienną losową o rozkładzie niezależnym od parametru $\theta \in \Theta$ oraz dla każdej realizacji $\boldsymbol{x}$ wektora losowego $\boldsymbol{X}$ odwzorowanie $\theta \to t(\boldsymbol{x}, \theta)$ jest monotoniczną funkcją parametru $\theta \in \Theta$. Stąd

1. można wyznaczyć takie dwie liczby $t_1, t_2 \in \mathbb{R}$, że

$$P_\theta\{t_1 < t(\boldsymbol{X}, \theta) < t_2\} \geq \delta \tag{8}$$

dla dowolnego $\theta \in \Theta$;
2. można dobrać takie dwie funkcje $\theta_L(\boldsymbol{x})$ i $\theta_R(\boldsymbol{x})$, dla których zachodzi równoważność

$$t_1 < t(\boldsymbol{x}, \theta) < t_2 \Leftrightarrow \theta \in (\theta_L(\boldsymbol{x}), \theta_R(\boldsymbol{x})). \tag{9}$$

Wtedy przedział $(\theta_L(\boldsymbol{X}), \theta_R(\boldsymbol{X}))$ jest przedziałem ufności dla parametru $\theta \in \Theta$, ponieważ

$$P_\theta\{\theta \in (\theta_L(\boldsymbol{x}), \theta_R(\boldsymbol{x}))\} = P_\theta\{t_1 < t(\boldsymbol{X}, \theta) < t_2\} \geq \delta \tag{10}$$

dla dowolnego $\theta \in \Theta$. Jeżeli znamy funkcję centralną, to konstrukcja przedziału ufności jest dość prosta. Trzeba jednak wiedzieć, jak taką funkcję skonstruować.

W tym przypadku wymagana jest dobra znajomość rozkładów prawdopodobieństwa zmiennych losowych.

Poza omówionymi wyżej kryteriami są jeszcze inne, związane z takimi pojęciami, jak *jednostajnie najdokładniejsze zbiory ufności*, *nieobciążone zbiory ufności* oraz *ekwiwariantne zbiory ufności*. Ich konstrukcja wymaga znajomości teorii weryfikacji hipotez. Wiele informacji na temat tych koncepcji można znaleźć w książce Lehmanna (1968) lub Bartoszewicza (1996).

5. Przykłady

Zakładamy, że próba $X_1, X_2, \ldots, X_n$ składa się z niezależnych zmiennych losowych o tym samym rozkładzie normalnym $N(\mu, \sigma^2)$. Przy powyższym założeniu, przyjmując oznaczenie

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \tag{11}$$

można udowodnić, że:

1. zmienna losowa $\frac{(n-1)S^2}{\sigma^2}$ ma rozkład chi-kwadrat z $n-1$ stopniami swobody;
2. zmienna losowa $\frac{\bar{X}-\mu}{S}\sqrt{n}$ ma rozkład *t*-Studenta o $n-1$ stopniach swobody.

Z podanych faktów wynika, że przy oznaczeniu $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ funkcje

$$t_1(\boldsymbol{X}, \mu) = \frac{\bar{X}-\mu}{S}\sqrt{n} \text{ oraz } t_2(\boldsymbol{X}, \sigma^2) = \frac{(n-1)S^2}{\sigma^2} \tag{12}$$

są funkcjami centralnymi. Pierwsza jest monotoniczna ze względu na parametr μ, a druga – ze względu na σ^2. Obie też mają rozkłady niezależne od parametrów μ i σ^2. Wektor $\theta = (\mu, \sigma^2)$ można zatem traktować jako parametr wielowymiarowy. Zwróćmy uwagę na to, że do tego miejsca parametr był jednowymiarowy. Przejście na dwuwymiarowość nie zmienia istoty rzeczy, co zostanie pokazane w przykładach. Niemniej, żeby uniknąć nieporozumień, zostanie przytoczona definicja przedziału ufności dla przypadku, gdy $\theta = (\theta_1, \theta_2)$.

Definicja. Przedział $(\theta_L(\boldsymbol{X}), \theta_R(\boldsymbol{X}))$, taki że

$$P_\theta\{\theta_1 \in (\theta_L(\boldsymbol{X}), \theta_R(\boldsymbol{X}))\} \geq \delta, \forall\theta \in \Theta, \tag{13}$$

nazywamy przedziałem ufności dla parametru θ_1, na poziomie ufności $\delta \in (0, 1)$.

Analogicznie definiuje się przedział ufności dla θ_2. W obu przypadkach przestrzeń parametrów Θ jest podzbiorem przestrzeni $\mathbb{R}^2$. W sytuacji, w której przedział ufności jest wyznaczany dla parametru θ_1, parametr θ_2 nazywa się *parametrem zaburzającym* lub *zakłócającym*. I na odwrót, gdy wyznaczamy przedział ufności dla θ_2, parametrem zakłócającym jest θ_1. O znaczeniu takich parametrów w estymacji lub weryfikacji hipotez statystycznych można przeczytać np. w pracach Lehmanna (1968, 1991).

Korzystając z funkcji centralnej $t_1(\boldsymbol{X}, \mu)$, budujemy przedział ufności dla parametru μ. Niech t_ν oznacza kwantyl rzędu ν rozkładu t-Studenta o $n-1$ stopniach swobody. Dla ustalonego $\delta_1 \in [0, 1]$ mamy

$$P_{\mu,\sigma^2}\{t_{\delta_1} < t_1(\boldsymbol{X}, \mu) < t_{\delta+\delta_1}\} = \delta, \forall(\mu \in R, \sigma^2 \in R_+). \tag{14}$$

Ponieważ

$$t_{\delta_1} < t_1(\boldsymbol{X}, \mu) < t_{\delta+\delta_1} \Leftrightarrow \mu \in \left(\bar{X} - t_{\delta+\delta_1}\frac{S}{\sqrt{n}}, \bar{X} - t_{\delta_1}\frac{S}{\sqrt{n}}\right), \tag{15}$$

przedział ufności dla wartości oczekiwanej μ na poziomie ufności δ ma postać

$$\left(\bar{X} - t_{\delta+\delta_1}\frac{S}{\sqrt{n}}, \bar{X} - t_{\delta_1}\frac{S}{\sqrt{n}}\right). \tag{16}$$

Stąd wynika również, że dla prawdopodobieństwa przeszacowania α_p zachodzi

$$\alpha_p = P_{\mu,\sigma^2}\left\{\mu \leq \bar{X} - t_{\delta+\delta_1}\frac{S}{\sqrt{n}}\right\} = 1 - (\delta + \delta_1), \tag{17}$$

a dla niedoszacowania α_n –

$$\alpha_n = P_{\mu,\sigma^2}\left\{\mu \geq \bar{X} - t_{\delta_1}\frac{S}{\sqrt{n}}\right\} = \delta_1. \tag{18}$$

Zauważmy, że w rozważanym przypadku, w którym σ^2 jest parametrem zaburzającym, mamy

$$A_\mu = \{\boldsymbol{x}: t_{\delta_1} < t_1(\boldsymbol{x}, \mu) < t_{\delta+\delta_1}\} \tag{19}$$

oraz

$$C(\boldsymbol{x}) = \{\mu: \boldsymbol{x} \in A_\mu\} = \left(\bar{x} - t_{\delta+\delta_1}\frac{S}{\sqrt{n}}, \bar{x} - t_{\delta_1}\frac{S}{\sqrt{n}}\right). \tag{20}$$

Ponieważ rozkład *t*-Studenta jest symetryczny względem 0, mamy $t_{\delta_1} = -t_{1-\delta_1}$ i przedział ufności dla wartości średniej można zapisać w postaci

$$\left(\bar{X} - t_{\delta+\delta_1}\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\delta_1}\frac{S}{\sqrt{n}}\right). \tag{21}$$

Przy symetrycznym podziale ryzyka przyjmujemy $\delta_1 = \frac{1-\delta}{2}$, ponieważ wtedy

$$\alpha_n = \alpha_p = \frac{1-\delta}{2}. \tag{22}$$

Zastępując δ przez $1 - \alpha$, otrzymamy znaną z podręczników postać

$$\left(\bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n}}\right). \tag{23}$$

Jeżeli przyjmiemy prawdopodobieństwo przeszacowania $1 - (\delta + \delta_1)$ lub niedoszacowania δ_1 równe 0, otrzymamy jednostronny przedział ufności

$$\left(-\infty, \bar{X} + t_\delta\frac{S}{\sqrt{n}}\right) \text{lub} \left(\bar{X} - t_\delta\frac{S}{\sqrt{n}}, \infty\right). \tag{24}$$

Jednostronne przedziały nie są z pewnością najwęższe, ponieważ ich szerokość jest nieograniczona. Aby znaleźć najwęższy przedział ufności, za δ_1 należy przyjąć taką wartość z przedziału $(0, 1 - \delta)$, dla której szerokość przedziału

$$\left(t_{\delta+\delta_1} - t_{\delta_1}\right)\frac{S}{\sqrt{n}} \tag{25}$$

jest najmniejsza. Osiągamy to poprzez minimalizację funkcji

$$l(\delta_1) = t_{\delta+\delta_1} - t_{\delta_1} = F^{-1}(\delta + \delta_1) - F^{-1}(\delta_1), \tag{26}$$

gdzie F oznacza dystrybuantę rozkładu *t*-Studenta o $n - 1$ stopniach swobody.

Dystrybuanta F oraz jej odwrotność F^{-1} są funkcjami różniczkowalnymi, a zatem

$$\frac{dl(\delta_1)}{d\delta_1} = \frac{1}{f(F^{-1}(\delta+\delta_1))} - \frac{1}{f(F^{-1}(\delta_1))}, \tag{27}$$

gdzie f jest funkcją gęstości rozkładu t-Studenta.

Pochodna funkcji $l(\delta_1)$ zeruje się, gdy

$$F^{-1}(\delta+\delta_1) = -F^{-1}(\delta_1). \tag{28}$$

Z własności kwantyli rozkładu t-Studenta wynika, że jest to możliwe tylko wtedy, gdy $\delta_1 = \frac{1-\delta}{2}$. Oznacza to, że najwęższy przedział ufności dla wartości średniej jest jednocześnie przedziałem o symetrycznym podziale ryzyka błędu. Jest również przedziałem o najmniejszej średniej szerokości tego przedziału, która w tym przypadku wynosi

$$l_n\left(\frac{1-\delta}{2}\right) E_{\mu,\sigma^2}\left(\frac{S}{\sqrt{n}}\right) = \sqrt{2}\,\frac{l_n\left(\frac{1-\delta}{2}\right)\Gamma\left(\frac{n}{2}\right)}{\sqrt{n(n-1)}\,\Gamma\left(\frac{n-1}{2}\right)}\,\sigma, \tag{29}$$

gdzie indeks n dopisany do wcześniej zdefiniowanej funkcji l służy podkreśleniu, że funkcja ta zależy od rozmiaru próby poprzez stopnie swobody rozkładu t-Studenta.

Należy zauważyć, że szerokość przedziału ufności jest proporcjonalna do odchylenia standardowego S z próby, a średnia szerokość owego przedziału – do parametru σ. Stąd też, aby kontrolować szerokość przedziału ufności, należy rozważyć szerokość względną (szerokość przedziału podzieloną przez odchylenie standardowe σ). W szczególności szukamy takiej minimalnej wielkości próby, by dla zadanego $d > 0$ zachodził jeden z warunków:
1. średnia względna szerokość przedziału nie przekracza d:

$$\sqrt{2}\,\frac{l_n\left(\frac{1-\delta}{2}\right)\Gamma\left(\frac{n}{2}\right)}{\sqrt{n(n-1)}\,\Gamma\left(\frac{n-1}{2}\right)} \le d, \tag{30}$$

2. z bardzo dużym prawdopodobieństwem względna szerokość przedziału nie przekracza d:

$$P\left\{l_n\left(\frac{1-\delta}{2}\right)\frac{S/\sigma}{\sqrt{n}} \leq d\right\} = \gamma, \tag{31}$$

dla ustalonego $\gamma \in [0,1]$ bliskiego 1.

Ponieważ zmienna losowa $\sqrt{n-1}\,S/\sigma$ ma rozkład chi z $n-1$ stopniami swobody, drugi warunek można zapisać w postaci

$$l_n\left(\frac{1-\delta}{2}\right)\frac{G_n^{-1}(\gamma)}{\sqrt{(n-1)n}} \leq d, \tag{32}$$

gdzie $G_n^{-1}(\gamma)$ oznacza kwantyl rzędu γ rozkładu chi z $n-1$ stopniami swobody.

Oba warunki można rozwiązać numerycznie ze względu na n. W tabl. 1 podano rozmiary próby dla zadanej dokładności d. W drugiej kolumnie znajdują się rozmiary dla szerokości oczekiwanej, a w trzeciej – dla szerokości, która jest realizowana z prawdopodobieństwem $\gamma = 0{,}9$. W obu przypadkach przyjęto poziom ufności $\delta = 0{,}95$.

Tabl. 1. Rozmiary próby dla zadanej dokładności d

d	Rozmiar próby dla szerokości	
	oczekiwanej	prawdopodobnej
0,45	78	93
0,40	98	116
0,35	128	148
0,30	173	196
0,25	248	276

Źródło: opracowanie własne.

Na przykład dla próby rozmiaru 98 oczekiwana względna szerokość przedziału ufności nie przekracza 40% odchylenia standardowego. Przy tej samej precyzji w drugim przypadku rozmiar próby jest większy. Możemy wtedy powiedzieć, że przy próbie rozmiaru 116 z prawdopodobieństwem 0,9 uzyskujemy przedział ufności o względnej szerokości nieprzekraczającej 40% odchylenia standardowego.

Podobnie można skonstruować przedziały ufności dla wariancji σ^2, przy czym w tym przypadku to μ jest parametrem zaburzającym. Korzystając z funkcji centralnej $t_2(\boldsymbol{x}, \sigma^2)$, otrzymujemy

$$P_{\mu,\sigma^2}\{\chi^2_{\delta_1} \leq t_2(\boldsymbol{X},\sigma^2) \leq \chi^2_{\delta+\delta_1}\} = \delta, \forall(\mu \in R, \sigma^2 \in R_+), \tag{33}$$

gdzie χ^2_ν oznacza kwantyl rzędu ν rozkładu chi-kwadrat z $n-1$ stopniami swobody oraz $\delta_1 \in [0, 1-\delta]$ jest ustaloną dowolnie wartością.

Z równoważności

$$\chi^2_{\delta_1} \leq t_2(\boldsymbol{X},\sigma^2) \leq \chi^2_{\delta+\delta_1} \Leftrightarrow \sigma^2 \in \left(\frac{(n-1)S^2}{\chi^2_{\delta+\delta_1}}, \frac{(n-1)S^2}{\chi^2_{\delta_1}}\right), \tag{34}$$

wynika, że przedział ufności dla wariancji σ^2 na poziomie ufności δ ma postać

$$\left(\frac{(n-1)S^2}{\chi^2_{\delta+\delta_1}}, \frac{(n-1)S^2}{\chi^2_{\delta_1}}\right). \tag{35}$$

Stąd prawdopodobieństwo przeszacowania

$$\alpha_p = P_{\mu,\sigma^2}\left\{\sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\delta+\delta_1}}\right\} = 1-(\delta+\delta_1), \tag{36}$$

a niedoszacowania

$$\alpha_n = P_{\mu,\sigma^2}\left\{\sigma^2 \geq \frac{(n-1)S^2}{\chi^2_{\delta}}\right\} = \delta_1. \tag{37}$$

Analogicznie do przypadku oszacowania wartości średniej, dla $\delta_1 = \frac{1-\delta}{2}$ otrzymamy przedział ufności o symetrycznym podziale ryzyka błędu. W przypadku braku ryzyka przeszacowania otrzymamy przedział

$$\left(0, \frac{(n-1)S^2}{\chi^2_{1-\delta}}\right), \tag{38}$$

a przy braku ryzyka niedoszacowania – przedział

$$\left(\frac{(n-1)S^2}{\chi^2_{1+\delta}}, \infty\right). \tag{39}$$

Najwęższy przedział ufności otrzymuje się dla δ_1 wyznaczonego z równania

$$\frac{1}{f(F^{-1}(\delta+\delta_1))}-\frac{1}{f(F^{-1}(\delta_1))}=0, \tag{40}$$

gdzie F i f oznaczają odpowiednio dystrybuantę i funkcję gęstości rozkładu chi-kwadrat z $n-1$ stopniami swobody (dla $n \geq 2$).

Ponieważ rozkład chi-kwadrat nie jest symetryczny, warunek ten można rozwiązać jedynie numerycznie. W tabl. 2 zawarto szerokości przedziałów ufności o symetrycznym podziale ryzyka błędu oraz przedziałów najwęższych. W ostatniej kolumnie podano prawdopodobieństwa niedoszacowania.

Tabl. 2. Szerokości przedziałów ufności o symetrycznym podziale ryzyka błędu oraz przedziałów najwęższych

Rozmiar próby	Szerokość przedziału		δ_1
	o symetrycznym podziale błędu	najwęższego	
10	0,259	0,221	0,048
20	0,075	0,069	0,044
30	0,038	0,036	0,042
40	0,024	0,023	0,040
50	0,017	0,016	0,039
100	0,006	0,006	0,035

Źródło: opracowanie własne.

Zagadnienie dotyczące przedziałów ufności dla wariancji w odniesieniu do ich zadanej szerokości można rozpatrywać w ujęciu względnym, tzn. w jednostkach szacowanej wariancji. Na przykład warunek, że średnia względna szerokość przedziału dla wariancji ma nie przekroczyć wartości d, zapisujemy następująco:

$$E_{\mu,\sigma^2}\left\{\frac{(n-1)S^2}{\sigma^2}\left(\frac{1}{\chi^2_{\delta_1}}-\frac{1}{\chi^2_{\delta+\delta_1}}\right)\right\}=\left(\frac{1}{\chi^2_{\delta_1}}-\frac{1}{\chi^2_{\delta+\delta_1}}\right)(n-1)\leq d. \tag{41}$$

Natomiast warunek, że z bardzo dużym prawdopodobieństwem (dla $\gamma \in (0,1)$ bliskiego 1) względna szerokość przedziału nie przekracza d, zapisujemy w postaci $P_{\mu,\sigma^2}\left\{\frac{(n-1)S^2}{\sigma^2}\left(\frac{1}{\chi^2_{\delta_1}}-\frac{1}{\chi^2_{\delta+\delta_1}}\right)\right\}\geq\gamma$ lub w postaci równoważnej:

$$\left(\frac{1}{\chi^2_{\delta_1}}-\frac{1}{\chi^2_{\delta+\delta_1}}\right)G_n^{-1}(\gamma)\leq d, \tag{42}$$

gdzie $G_n^{-1}(\gamma)$ oznacza kwantyl rzędu γ rozkładu chi-kwadrat z $n-1$ stopniami swobody.

Oba podane warunki można rozwiązać numerycznie ze względu na rozmiar próby n. Celem jest znalezienie takiego najmniejszego n, dla którego zachodzi pierwszy albo drugi warunek. Należy przy tym pamiętać, że kwantyle $\chi^2_{\delta_1}$ i $\chi^2_{\delta+\delta_1}$ również zależą od rozmiaru próby. Prawdopodobieństwo niedoszacowania δ_1 można uprzednio ustalić lub przyjąć takie, które daje najwęższy przedział ufności. W tym drugim przypadku prawdopodobieństwo to również zależy od rozmiaru próby. Na przykład dla $\delta_1 = \frac{1-\delta}{2}, \delta = 0{,}95$ i $\gamma = 0{,}9$ w tabl. 3 podano wielkości próby dla zadanej szerokości względnej przedziału ufności.

Tabl. 3. Rozmiary próby dla oczekiwanej oraz prawdopodobnej szerokości przedziału ufności

d	Rozmiar próby dla szerokości	
	oczekiwanej	prawdopodobnej
0,45	161	203
0,40	201	248
0,35	260	314
0,30	350	415
0,25	501	578

Źródło: opracowanie własne.

Na przykład dla próby rozmiaru 201 oczekiwana względna szerokość przedziału ufności nie przekracza 40% szacowanej wariancji. Natomiast aby względna szerokość z prawdopodobieństwem 0,9 również nie przekroczyła 40% szacowanej wariancji, rozmiar próby musi być większy i wynosić co najmniej 248. W obu przypadkach prawdopodobieństwo niedoszacowania jest równe prawdopodobieństwu przeszacowania i wynosi 2,5%.

6. Wymiar edukacyjny

Podręczniki oraz skrypty do statystyki dla studentów zawierają wiedzę albo o wysokim stopniu ogólności, jak prace Lehmanna (1968) lub Bartoszewicza (1996), albo przeciwnie – o niewielkim wycinku wnioskowania statystycznego. W obu przypadkach brakuje odniesień do – przedstawionych w niniejszym artykule – koncepcji podziału ryzyka błędu przeszacowania lub niedoszacowania parametru oraz koncepcji optymalnej liczebności próby.

Pierwsza z wymienionych koncepcji jest użyteczna, gdy niesymetryczny podział ryzyka błędu jest lepszy od symetrycznego – wówczas gdy strata wynikająca z niedoszacowania danego parametru jest inna niż wynikająca z przeszacowania. Na przykład lepiej nie doszacować niż przeszacować wartość zagrożoną, która zabezpiecza banki przed utratą płynności finansowej.

Druga koncepcja poszerza, ujęte w programach nauczania, zagadnienie doboru wielkości próby. Standardowe podejście wymaga oszacowania odchylenia standar-

dowego przed ustaleniem rozmiaru próby, co wymusza pobranie próby wstępnej albo korzystanie z wyników innych badań. Działania te mogą okazać się niepraktyczne, a otrzymane oszacowania – nieprecyzyjne. Ponadto dotyczą jedynie przypadku, w którym podział ryzyka błędu pomiędzy przeszacowanie a niedoszacowanie jest symetryczny. Są to wady, których nie ma zaprezentowany w artykule dobór wielkości próby.

Wyznaczenie przedziałów ufności o niesymetrycznym podziale ryzyka błędu oraz dobór optymalnych wielkości nie są skomplikowane. Wartości kwantyli znanych rozkładów prawdopodobieństwa oraz funkcji gamma można wyznaczać w arkuszach kalkulacyjnych oraz w ogólnie dostępnych pakietach statystycznych. Iloraz funkcji gamma w nierówności (30) może powodować problemy numeryczne. Jeżeli jednak iloraz ten będzie odpowiednio liczony (np. przy wykorzystaniu własności rekurencyjnej funkcji gamma), problemy te nie wystąpią.

Należy zwrócić uwagę, że przedstawione przykłady dotyczą szczególnego przypadku rozkładu normalnego, który jest rozkładem typu ciągłego. Konstrukcja przedziałów ufności dla parametrów rozkładu dyskretnego okazuje się pod pewnymi względami trudniejsza. Z tego powodu w edukacji statystycznej popularne stały się asymptotyczne przedziały ufności, np. dla wskaźnika struktury (rozumianego jako odsetek obiektów całej populacji mających interesującą nas własność). Przedziały te nie są jednak przedziałami ufności w myśl podanej definicji i można je z powodzeniem zastąpić przedziałem dokładnym (Zieliński, 2009), a nawet, dokładając pewien mechanizm losowy, poprawić jego właściwości (Zieliński, 2017).

7. Podsumowanie

W artykule przedstawiono przykłady konstrukcji przedziałów ufności na podstawie kryteriów wynikających z kontroli ryzyka przeszacowania i niedoszacowania parametru oraz zagadnienia doboru wielkości próby gwarantującej zadaną precyzję oszacowania. Podane kryteria uwzględniają ewentualny wymóg asymetrii tego ryzyka oraz, w przypadku najkorzystniejszego wyboru optymalnego rozmiaru próby, dają rozwiązania niezależne od szacunkowych wartości nieznanych parametrów. W ten sposób podane konstrukcje mogą, w ramach typowej edukacji statystycznej, poszerzyć zakres zagadnień związanych z konstrukcją przedziałów ufności, a także przyczynić się do ich lepszego zrozumienia.

Bibliografia

Bartoszewicz, J. (1996). *Wykłady ze statystyki matematycznej*. Państwowe Wydawnictwo Naukowe.

Kotz, S., Johnson, N. L. (red.). (1992). *Breakthroughs in Statistics: Volume 2. Methodology and Distribution*. Springer-Verlag. https://doi.org/10.1007/978-1-4612-4380-9.

Krzyśko, M. (1996). *Statystyka matematyczna*. Państwowe Wydawnictwo Naukowe.

Krzyśko, M., Adamczewski, W., Berger, J., Gołata, E., Kruszka, K., Łazowska, B. (2018). *Statystycy polscy. Biogramy*. Główny Urząd Statystyczny. https://bws.stat.gov.pl/BWS/Archiwum/gus_bws_67_Statystycy_polscy_Biogramy.pdf.

Lehmann, E. L. (1968). *Weryfikacja hipotez statystycznych*. Państwowe Wydawnictwo Naukowe.

Lehmann, E. L. (1991). *Teoria estymacji punktowej*. Państwowe Wydawnictwo Naukowe.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*(4), 558–625. https://doi.org/10.2307/2342192.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London: Series A. Mathematical and Physical Sciences*, *236*(767), 333–380. https://doi.org/10.1098/rsta.1937.0005.

Zieliński, R. (2009). Przedział ufności dla frakcji. *Matematyka Stosowana*, *9*(50), 76–90.

Zieliński, W. (2017). The shortest Clopper-Pearson randomized confidence interval for binomial probability. *REVSTAT – Statistical Journal*, *15*(1), 141–153.

Wiadomości Statystyczne. The Polish Statistician, 2022, vol. 67, 2, 36–50 DOI: 10.5604/01.3001.0015.7719
Z dziejów statystyki / From the history of statistics
Zgłoszony/submitted: 30.08.2021, poprawiony/revised: 28.12.2021, zaakceptowany/accepted: 19.01.2022

The 1897 Census in the Kingdom of Poland[1]

Piotr Rachwał[a]

Abstract. The aim of the article is to present the issues relating to the organisation, conduct and processing of the results of the first population census in the Kingdom of Poland in 1897 in a concise way. Although the text focuses on the lands of the Kingdom of Poland which were under the Russian rule the aforementioned issues have been discussed in a broader context of the activity of the central institutions in St. Petersburg. Polish historiography addresses these matters only to a limited extent and very few authors have dealt primarily with the notion of the completeness of census data.

The provincial authorities and the Warsaw Statistical Committee played a major role in the preparation and conduct of the census in the Kingdom of Poland. It should be emphasised, however, that the greatest effort related to data collection rested with the lowest-level local government. At the stage of the central preparation of the census results, modern technologies, i.e. Hollerith's calculating machines, were used. The data that was collected on forms containing 14 questions were coded on special perforated cards, which significantly improved the performance of the relevant calculations. The level at which the census was carried out and its final outcome are assessed relatively high. On the other hand, there are certain reservations as to the credibility of the census results relating to the Kingdom of Poland, the most serious of which concern the reporting of the age, the professional structure, as well as the religion and mother tongue of the respondents.

Keywords: 1897 Census, Kingdom of Poland, Warsaw Statistical Committee, perforated cards
JEL: B00

Powszechny spis ludności w Królestwie Polskim w 1897 roku

Streszczenie. Celem artykułu jest przedstawienie zagadnień związanych z organizacją, przebiegiem i opracowaniem wyników pierwszego powszechnego spisu ludności w Królestwie Polskim, który przeprowadzono w 1897 r. Uwagę skupiono na ziemiach Królestwa Polskiego, znajdującego się pod panowaniem rosyjskim, niemniej jednak starano się przedstawić problematykę spisu w szerszym kontekście działalności władz centralnych w Petersburgu. W polskiej historiografii zagadnienia te podejmowali nieliczni autorzy i jedynie w ograniczonym zakresie, przede wszystkim dotyczącym oceny kompletności danych spisowych.

[1] The article is based on a paper delivered at the *Past towards the future – censuses on Polish lands* (Pol. *Przeszłość ku przyszłości – powszechne spisy ludności na ziemiach polskich*) scientific seminar, organised by the Committee on Demographic Studies of the Polish Academy of Sciences in collaboration with the Polish National Historical Committee of the Polish Academy of Sciences, held on 10th March, 2021 in Warsaw.
Artykuł został opracowany na podstawie referatu wygłoszonego na seminarium naukowym *Przeszłość ku przyszłości – powszechne spisy ludności na ziemiach polskich*, zorganizowanym przez Komitet Nauk Demograficznych PAN we współpracy z Komitetem Nauk Historycznych PAN, które odbyło się 10 marca 2021 r. w Warszawie.

[a] Katolicki Uniwersytet Lubelski Jana Pawła II, Wydział Nauk Humanistycznych, Instytut Historii, Polska / The John Paul II Catholic University of Lublin, Faculty of Humanities, History Institute, Poland.
ORCID: https://orcid.org/0000-0003-1903-4115. E-mail: piotrach@kul.pl.

Główną rolę w przygotowaniu i przeprowadzeniu spisu w Królestwie Polskim odegrały władze gubernialne oraz Warszawski Komitet Statystyczny. Należy jednak podkreślić, że największy wysiłek związany z gromadzeniem danych spoczywał na władzach samorządowych najniższego szczebla. Na etapie centralnego opracowywania wyników spisu wykorzystano nowoczesne technologie, tj. maszyny liczące Holleritha. Dane zebrane na formularzach (które zawierały 14 pytań) kodowano na specjalnych kartach perforowanych, co znacząco usprawniło obliczenia. Wyniki spisu i sposób jego przeprowadzenia należy ocenić stosunkowo dobrze. W przypadku Królestwa Polskiego największe zastrzeżenia do wiarygodności spisu dotyczą raportowania wieku ludności, struktury zawodowej, a także wyznania i języka ojczystego.

Słowa kluczowe: powszechny spis ludności 1897 r., Królestwo Polskie, Warszawski Komitet Statystyczny, karty perforowane

1. Introduction

The collection of data on the condition and structure of the population stands a long tradition reaching back to as early as the ancient times, and the Roman Empire had a particularly well-developed state apparatus facilitating the process of gathering information (Suder, 2003). Reporting of this type weakened in the Middle Ages, and it was not until the Modern Era that the knowledge of the demographics of individual administrative units became important to both church and state authorities. What all historical censuses had in common was the purpose for which they were conducted, i.e. to allow the efficient management of the state, which involved acquiring funds effectively for its functioning and determining the country's mobilisation possibilities, i.e. its military potential (Domschke & Goyer, 1986; Goyer & Domschke, 1983; Goyer & Draaijer, 1992; Łukasiewicz, 2009).

However, the modern concept of a population census[2] – in terms of covering all the inhabitants of a given area, regardless of age, sex, professional or state affiliation, etc. – was not developed until the 18th century (Holzer, 1999). The first country to have conducted a census according to the modern criteria was Iceland. The census took place in 1703 in the form of a personal census. Another census was conducted in Sweden in 1749 and since then censuses have been conducted there every five years. Censuses were also organised in the United States of America in 1790, in France in 1800, and in England in 1801, and take place every 10 years (Thorvaldsen, 2018).

In the Polish territory, the first general censuses were conducted in the 19th century. The Polish state was then partitioned, therefore the statistical apparatuses of the individual partitioning states were responsible for organising the censuses. Two of them were held in the Duchy of Warsaw, which was granted considerable autonomy. Both were nominal; the first lasted from July 1807 to the end of 1808 and the second

[2] A population census is a systematic recording of information on all members of a population usually residing in a country (*de jure*) or present at the time of the enumeration (*de facto*). The most important features of censuses are: financing by the national government, agreement on areas to be included, universality, individual enumeration, simultaneity of enumeration, periodicity, and publication and dissemination of the results (Bailar, 2001; Kreyenfeld & Willekens, 2015).

from October 1809 to March 1810 (Grossman, 1925). In Prussia, the first nominal census was held in 1840, and in the Austrian Empire in 1857 (Łukasiewicz, 2009).[3]

The aim of the article is to present the issues relating to the organisation, conduct and processing of the results of the first population census in the Kingdom of Poland in 1897 in a concise way. The Kingdom of Poland, established at the Congress of Vienna in 1815, connected by a personal union with Russia, was strictly dependent on and controlled by it until Poland regained independence in 1918. The census of 1897 was the only one that took place in the Russian Empire before World War I.[4] It covered the entire Empire (excluding the Grand Duchy of Finland) and was supervised by the central institutions in St. Petersburg, thus the issues raised in this paper are considered in a broader context – not only in relation to the Kingdom of Poland, but also to the activity of the central administration. The article discusses the preparations for the census, its organisation and conduct, as well as the work related to the processing of the results. The purpose of the article does not include the analysis of the statistical data collected during the census (which can be found in Janczak, 1982; Nietyksza, 1971; Pruss, 1977; Szulc, 1920).

2. The origin and technical organisation of the census

In tsarist Russia, in the second half of the 19th century, a new socio-economic reality was forming as a result of intensified industrialisation, urbanisation and, above all, the Emancipation Reform, which came into force in 1861. The reform granted peasants various rights, including personal freedom, the possibility of social advancement or greater mobility, albeit to a limited extent (Krnilov, 1905; Litvak, 1991). The Emancipation Reform was introduced in the Kingdom of Poland in 1864 (Bardach et al., 1994). What significantly prompted the decision of the authorities in St. Petersburg to conduct a census which would provide detailed data on the country's demography was, in addition to the aforementioned social transformations, the accumulation of problems hindering the efficient management of the

[3] Successive censuses of the population of the Polish territories under Austrian and Prussian rule took place in Galicia in 1869, 1880, 1890, 1900 and 1910, and in the Prussian partition in 1843 and 1846. The latter was the year that the principle of a one-day census organised every 3 years was adopted, and since 1875 censuses were carried out every 5 years. In the Russian Empire, the second census was scheduled for 1910, but it was postponed until 1915, yet even then it did not take place. Thus, the 1897 Census was the first and only statistical survey of this type conducted in tsarist Russia. A separate issue here is the comparability of the collected and processed census data in terms of e.g. grouping the collected statistical data, the statistical categories used, etc. Comparative studies in this area are yet to be conducted. However, some older studies in this area, including those of Krzyżanowski and Kumaniecki (1915), are worth mentioning here. According to Łukasiewicz's general opinion, the results of the censuses carried out in Poland by the authorities of the partitioning states were virtually incomparable due to the chronological differences and the use of varying statistical methods and terminology. It was only the censuses carried out in the Second Polish Republic that allowed a full comparability of the results.

[4] It does not mean, however, that no official population statistics were kept at that time. Activities in this area rested with the governorate authorities, and the local press also played a major role in this respect since its representatives were often members of special committees formed for this purpose (Młodak, 2011).

state, including the effective coordination and response to all kinds of crises. Serious negligence in this regard was exposed by the famine that hit Russia in the years 1891–1892. The crop failure covered an area of over 1.4 million square kilometres, mainly in the central part of the Empire and lands in the Volga basin, including the provinces: Nizhny Novgorod, Kazan, Ryazan, Penza, Tula, Simbirsk, Saratov, Samara and Tambov. An estimated 14 to 20 million people were starving at that time, of whom nearly 400 thousand died, mostly from various diseases that attacked their exhausted by hunger bodies (Bazylow, 1970; Robbins, 1975).[5]

A draft census that would cover the entire territory of the Empire[6] (Clem, 2016) had already been discussed at the 1st All-Russian Congress of Statisticians (1870), and then again in 1876 at the 8th Session of the International Statistical Congress. In February 1877, the draft regulations of the census developed by a commission at the Ministry of Finance were submitted to the State Council, yet due to some dignitaries' opposition and a tense political situation (caused by the war with Turkey in the years 1877–1878), the plans were not implemented at that time (Nietyksza, 1971; Schwartz, 1986).[7]

The main initiators of the census and its long-standing promoters were Pëtr Petrovič Semënov-Tân-Šanskij, geographer and statistician, Chief of the Central Statistical Bureau in the years 1864–1875, and his close associate, Nikolaj Aleksandrovič Trojnickij, Chairman of the Council for Statistics at the Russian Ministry of the Interior (Edlund, 1999). Eventually, the intentions and plans were accepted on 17th June 1895; they were then approved by the authorities and preparatory work began. In July 1896 the date of the census was set for 28th January 1897 in the Old Style, i.e. 9th February 1897 according to the Gregorian calendar.

The body supervising the preparation and conduct of the census was the Central Census Committee in St. Petersburg, established especially for this purpose. Formally, it was subordinate to the Ministry of the Interior. Committees in governorates and counties were lower level bodies subordinate to this committee. In some larger cities, separate committees were established regardless of the county ones. It was the county committees that were responsible for dividing the county or city into districts. Each of these districts included approximately 25 thousand inhabitants in the cities, and usually one municipality in villages. The districts were

[5] According to observers and commentators of public life, such serious consequences of the famine could have been avoided if the central authorities had effectively collaborated with the local administration. The crisis highlighted the widespread corruption, ineffectiveness of the rulers and, most importantly, the central government's lack of awareness of the real situation in the provinces.

[6] By the end of the 19th century, Russia had an area of over 22 million square kilometres and in this respect, next to the British Empire, it was the vastest state in the world. Russia covered about one-sixth of the world's land.

[7] In the years 1860–1889, 79 censuses were conducted regionally and in large urban centres in the territory of the Russian Empire: St. Petersburg (1864, 1869, 1881, 1890), Moscow (1871, 1882), Tver (1869), Vladimir, Kazan and Kyiv (1874). In the Kingdom of Poland, the census took place in the Radom Governorate in 1868, in the Suwałki Governorate in 1889 and in the Łomża Governorate in 1890. In Warsaw such a census was conducted in 1882.

headed by managers who selected census enumerators, supervised their training and controlled their activity.

It is worth noting that in the second half of the 19th century in the Kingdom of Poland, institutions dealing with population statistics were decentralised. In January 1867, the Kingdom of Poland was divided into 10 governorates with seats in Kalisz, Kielce, Lublin, Łomża, Piotrków, Płock, Radom, Siedlce, Suwałki and Warsaw. In mid-1868, the Government Commission for Internal Affairs, Clergy and Public Education – the institution responsible for statistical reporting related to various aspects of the state's functioning – was dissolved. From then on, these responsibilities were shifted onto the editors of the *Obzor* (*Review*) journals, published annually in each governorate, providing current data on the condition, structure and natural movement of the population (Berger, 2002, 2017; Bornstein, 1920).

The Warsaw Statistical Committee played a significant role in statistical reporting in the Kingdom of Poland at that time. Chaired by Professor G. F. Simonenko, the institution took an active part in the 1897 Census. The Committee's leadership intended to assume as many competences related to the organisation and conduct of the census as possible, but the authorities in St. Petersburg did not agree to it. What is important here is that the initiatives undertaken by the Warsaw Statistical Committee were pro-Russian and aimed to spread relevant propaganda, often at the expense of the deviation from the methodological rules of statistical surveys.[8]

The obligation to collect data rested with census enumerators. Each of them had about 400 households (about 2 thousand inhabitants) to register in the countryside and about 150 apartments (750 people) in the cities. In the Kingdom of Poland, nearly 15 thousand census enumerators were appointed. According to the guidelines, anyone who wanted to become a census enumerator had to read and write well, be intelligent and familiar with the local relations, have high moral qualifications and the trust of the district manager and the local population. Before starting their work and after completing the training, census enumerators were to complete a number of census questionnaires as a test. In Warsaw, census enumerators were mainly students, whereas in provinces usually local officials and teachers (Szulc, 1920). All information was to be obtained by census enumerators through a personal interview; obtaining information from any other sources was forbidden. The headings of all the census sections and the instructions for the questionnaire were translated into local languages – 19 of them in total. The instructions were printed and translated in magazines for informational and educational purposes (Census in Rural Russia, 1897). The census enumerators started the first round 20–30 days before the main census in rural areas and 5–10 days prior in the cities. While rural residents were registered by the census

[8] This is especially visible with regard to the census registration of the population of the Greek Catholic denomination, where the undertaken actions affected the results of the census, e.g. in the case of some counties in the Siedlce Governorate.

enumerators in person, the residents of manor houses and towns completed the census questionnaires themselves. The census enumerators had to check each and every filled in form. The obtained information was verified on the day of the census, which began at dawn and ended the following day in the cities, and after a maximum of four days in rural areas (Szulc, 1920).

3. Census questionnaires

The census was conducted by household (a separate questionnaire for each household). However, the term 'household' was not precisely defined. The instructions for the city census enumerators stated that this was a separate apartment; for the ones in rural areas it was not specified at all. The census questions in the Kingdom of Poland were translated into Polish and German, and additionally into Lithuanian in the Suwałki Governorate. However, no consistency was shown in this regard, as, for example, the questionnaires in Warsaw were not translated, so they were available only in Russian. This discrepancy might have perhaps resulted from the organisers' ill will.[9] The instructions stated that the census questionnaire should include all people present in a given household, including not only permanent residents, but also those who stayed there temporarily, as well as permanent residents who were absent due to their official duties.[10] The census questionnaires had been prepared for five groups. The first questionnaire was intended for peasants who were subsistence farmers. The second one was for the population inhabiting estates. The third type was for city dwellers. A separate questionnaire had also been prepared for the military population (consisting of 12 questions), as well as for foreign students, the clergy and the charges of charitable organisations (Edlund, 1999). The basic questionnaire generally consisted of 14 questions concerning household members (Figure):

1. Family name and given name;
2. Sex (male, female);
3. Relation with regard to the head of the household and the head of the family;
4. Age (how many years or months have passed since birth);
5. Marital status (single, married, widowed, divorced);
6. Social class, rank or title;
7. Place of birth (if other than 'here', then which governorate, county, city);

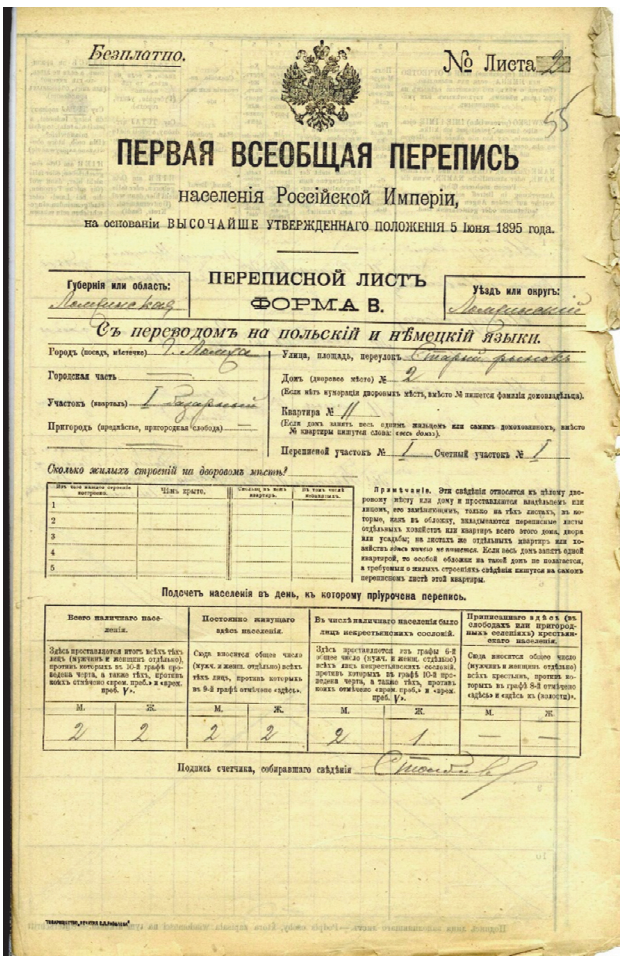
[9] This situation may have resulted from the lack of an appropriate number of prepared census cards in languages other than Russian, although the census itself showed that among the people who lived in Warsaw at that time over 59 thousand declared Russian as their mother tongue and over 421 thousand Polish. (Tsentral'nyy Statisticheskiy Komitet, 1904, vol. 51, p. 2 – the published census results are available in digital repertory of the Stefan Szulc Central Statistical Library: http://statlibr.stat.gov.pl/exlibris/aleph/a22_1/apache_media/887JXBFNTR12V6DX8HS8ASVNX9HMUV.pdf).

[10] Hence, the census combines two approaches to determining the state of the population, i.e. *de jure* population and *de facto* population. A detailed list in this respect was based on the already published census results – see: Craumer (1986). For the census instruction see: Główny Urząd Statystyczny (2002).

8. Place of registration. Are they registered 'here' in the population registry; if not, then where?;
9. Usual place of residence. Is it 'here'? If not, then where? (governorate, county, city);
10. Notice of absence, departure and temporary stay at the census location;
11. Religious denomination;
12. Native language;
13. Education (a. Are they literate?; b. Where did they go to school?);
14. Occupation, trade, industry, position of office or service (a. primary source of income; b1. secondary, b2. military service).

Figure. The 1897 Population Census Form from the Łomża Governorate, Łomża District

A. Sign. 10, p. 55



Безплатно. № Листа 1

55

ПЕРВАЯ ВСЕОБЩАЯ ПЕРЕПИСЬ

населенія Россійской Имперіи,

на основаніи ВЫСОЧАЙШЕ УТВЕРЖДЕННАГО ПОЛОЖЕНІЯ 5 іюня 1895 года.

Губернія или область: Ломжинская

ПЕРЕПИСНОЙ ЛИСТЪ

ФОРМА В.

Уѣздъ или округъ: Ломжинскій

Съ переводомъ на польскій и нѣмецкій языки.

Городъ (посадъ, мѣстечко) г. Ломжа

Городская часть —

Участокъ (квартал) 1 Городской

Пригородъ (предмѣстье, пригородная слобода) —

Улица, площадь, переулокъ Старой рынокъ

Домъ (дворовое мѣсто) № 2

(Если мѣсть нумерація дворовыхъ мѣстъ, вмѣсто № пишется фамилія домовладѣльца).

Квартира № 11

(Если домъ занятъ весь однимъ жильцомъ или самимъ домовладѣльцемъ, вмѣсто № квартиры пишутся слова: весь домъ).

Переписной участокъ № 1 Счетный участокъ № 1

Сколько жилыхъ строеній на дворовомъ мѣстѣ?

Изъ нихъ каменныхъ строеній выстроено.	Чѣмъ крыто.	Сколько въ нихъ квартиръ.	Въ томъ числѣ меблированныхъ.
1			
2			
3			
4			
5			

Примѣчаніе. Эти свѣдѣнія относятся къ цѣлому дворовому мѣсту или дому и приставленнои владѣльцемъ или лицомъ, его замѣщающимъ, только на тѣхъ листкахъ, изъ которыхъ, или въ обложку, вкладываются переписные листы отдѣльныхъ хозяйствъ или квартиръ всего этого дома, двора или усадьбы; на листкахъ же отдѣльныхъ квартиръ или хозяйствъ объ этомъ ничего не помѣщается. Если весь домъ занятъ одной квартирой, то особой обложки на такой домъ не полагается, а требуемые о такихъ строеніяхъ свѣдѣнія вносятся на самомъ переписномъ листѣ этой квартиры.

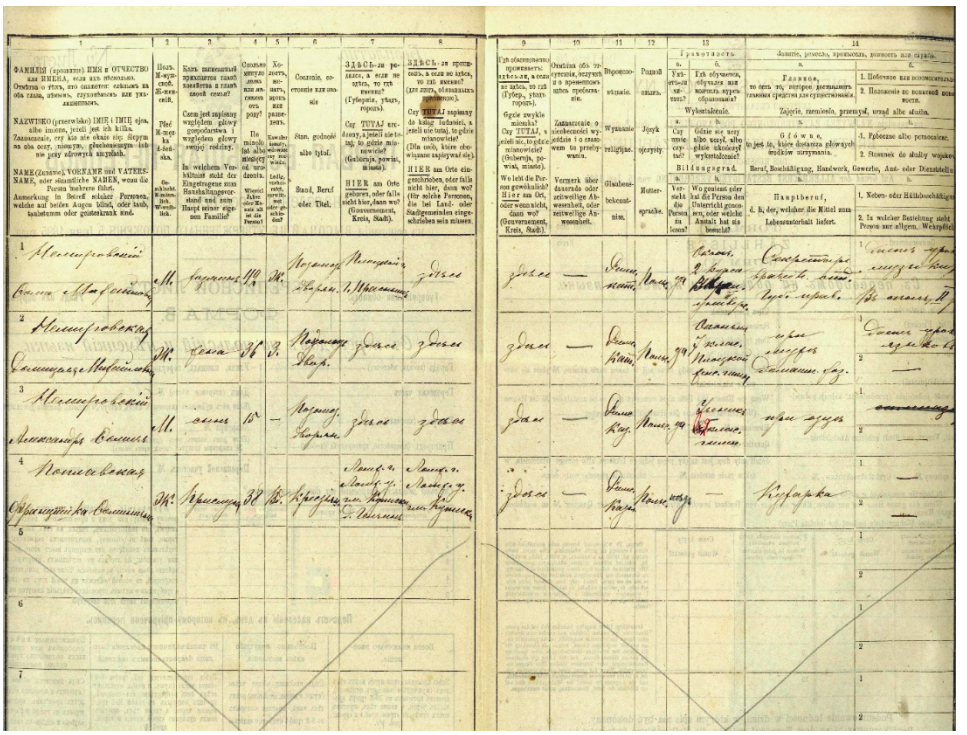
Подсчетъ населенія въ день, къ которому пріурочена перепись.

Всего наличнаго населенія.		Постоянно живущаго здѣсь населенія.		Въ числѣ наличнаго населенія было лицъ некрестьянскихъ сословій.		Приписаннаго вдѣсь (въ слободахъ или пригородахъ сельскаго) крестьянскаго населенія.	
Здѣсь проставляется итогъ всѣхъ тѣхъ лицъ (мужчинъ и женщинъ отдѣльно), противъ которыхъ въ 10-й графѣ проведена черта, а также тѣхъ, противъ коихъ отмѣчено «врем. проб.» и «врем. проб. у».		Сюда вносится общее число (муж. и женщ. отдѣльно) всѣхъ тѣхъ лицъ, противъ которыхъ въ 9-й графѣ отмѣчено «здѣсь».		Здѣсь проставляется изъ графы 9-й общее число (муж. и женщ. отдѣльно) всѣхъ тѣхъ некрестьянскихъ сословій, противъ которыхъ въ графѣ 10-й проведена черта, а также тѣхъ, противъ коихъ отмѣчено «врем. проб.» и «врем. проб. у».		Сюда вносится общее число (мужчинъ и женщинъ отдѣльно) всѣхъ крестьянъ, противъ которыхъ въ графѣ 8-й отмѣчено «здѣсь» и «здѣсь изъ (волости)».	
М.	Ж.	М.	Ж.	М.	Ж.	М.	Ж.
2	2	2	2	2	1	—	—

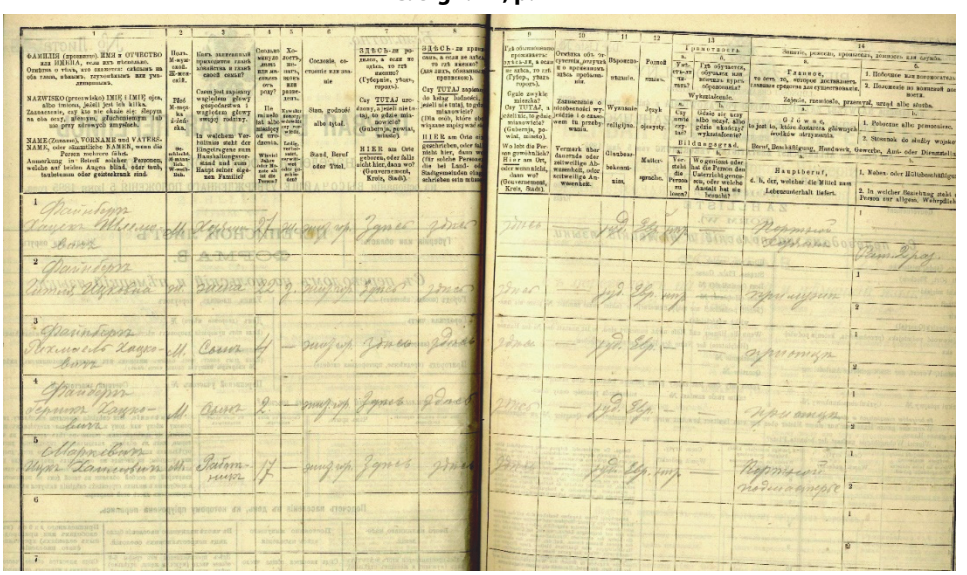
Подпись счетчика, собиравшаго свѣдѣнія Стецкій

Figure. The 1897 Population Census Form from the Łomża Governorate, Łomża District (cont.)

B. Sign. 10, p. 56



C. Sign. 11, p. 24



Source: Census Commission of the Łomża District 1896–1897, fond 19. The Łomża branch of the Białystok National Archives.

Although the census questions were clearly specified, certain problems with providing an unambiguous answer could arise in questions 9, 10 and 14, i.e. about the place of residence (stay) and occupation. Stefan Szulc,[11] when assessing the 1897 Census, noticed the ambiguity of the proposed responses, although in general his opinion of the questionnaire was positive. At the same time he emphasised the usefulness and significance of the data collected through these questionnaires: 'Despite all the reservations, it should be acknowledged that the questionnaire in the form given could be the basis for a work quite well done' (Szulc, 1920, pp. 5–6).

4. The processing and publication of the results

Having completed the collection of data, each census enumerator organised the material, calculated the population of his or her area on separate forms and submitted them to the district manager. After verifying them, the manager made copies of the forms. The originals were sent to St. Petersburg, while the copies were kept in the archives of the governorates. 'Representatives to arrange the census activities' were also appointed as an additional control body. Lieutenant-General S. I. Tolstoy was the representative in the Kingdom of Poland and his assistant was P. A. Baracz, who was one of the editors at the Warsaw Statistical Committee (Brûhanova & Ivanova, 2019; Szulc, 1920).

Next, the employees of the St. Petersburg Census Bureau transferred the information from the census questionnaires onto special perforated cards, individually for each interviewed person, and the data coded in such a way were then automatically analysed, providing the required statistical tabulations. The people who performed this task were called 'dietarjusze' and worked under the supervision and guidance of controllers and editors. The editors explained any arising doubts. The counting process itself was performed using Herman Hollerith's counting machines. The same technology had been used to process the data from the census conducted in the United States of America in 1890. It should be emphasised that it was an innovative technology and, as the organisers of the census stated, without it analysing such a vast amount of data in a reasonable amount of time would have been impossible. It is estimated that the entire project cost 7 million roubles (Bashe et al., 1985).[12]

[11] Stefan Aleksander Szulc (1881–1956) – Polish statistician, economist, the President of the Polish Statistical Association and the head of the Central Statistical Office.

[12] A meeting between Nikolai Alexandrovich Troynitsky, the Director of the Central Statistical Committee and Hollerith, the inventor of the punch card tabulating machine and founder of the Tabulating Machine Company (which in 1924 evolved into the IBM corporation) took place in St. Petersburg on 15th December 1896. Under their agreement, Hollerith's company leased Russia 35 counting machines, which had already been used in other censuses. The manufacturing of these devices was complicated and expensive, hence the decision to lease them. The necessary additional equipment, i.e. 70 tabulators with sorting machines and 500 punchers, was sold to Russia at a competitive price. The equipment was assembled in St. Petersburg under Hollerith's supervision. For further information, see: Anan'eva (n.d.).

The publication of the preliminary results of the census began in the same year, although the concept of the main publication was developed several years later. The published results provided data on the size of the actually present population as well as on the total population permanently residing in the area covered by the census, on those temporarily present and absent. The presentation in the tables was territory-based, with a varying degree of detail of the gathered information. The processed data was published in a monumental publication which consisted of 89 volumes (Tsentral'nyy Statisticheskiy Komitet, 1904). Data from each governorate was published in a separate volume, each consisting of an introduction and 28 tables. The data was printed at the county and city level. The data relating to the Kingdom of Poland was printed in volumes LI–LX and published in 1904 (volume LI was published separately for the Warsaw Governorate and Warsaw itself). The published tables contained the following information:

I. Actual and permanent population;
II. Distribution of the population by household and the composition of these households;
III. a. Population by ten-year age group;
b. Distribution by sex, age (in years) and literacy;
IV. Children below the age of one by number of months lived;
V. Distribution of the population by marital status and age group;
VI. Distribution of the population by social class group and place of birth;
VII. Distribution of the population born outside the census area by place of birth;
VIII. Distribution of the population by social class;
IX. Distribution of the population by literacy and education, as well as by social class and age group;
X. Distribution of the population by social class group and marital status;
XI. Distribution of foreign subjects by country;
XII. Distribution of the population by religious denomination;
XIII. Distribution of the population by native language;
XIV. Distribution of the population by religious denomination and native language;
XV. Distribution of the population by native language, literacy and age group;
XVI. Distribution of the population by marital status and native language;
XVII. Distribution of the population with physical disabilities by age group;
XVIII. Distribution of the population with physical disabilities by native language;
XIX. Distribution of the population with physical disabilities by social class group;
XX. Distribution of the population by occupation and age group;
XXI. Distribution by occupation group;
XXII. Distribution of occupation groups and nationality by native language;

XXIII. Distribution of the population active in sedentary farming, nomadic farming and herding, fishing and hunting by side activity;
XXIV. Distribution of the population by native language and social class;
XXV. Distribution of the population by religious denomination and ten-year age group.

Making compilations of data covering larger areas was a cumbersome and difficult process. Compilations relating to the Kingdom of Poland were presented in some tables of the *General Tabulation* (*Zestawienie ogólne…*, 1905). The composition of the Russian Empire's population in terms of occupation was additionally compiled in a separate publication. In total, about 400 different occupations were distinguished and the information about the Polish Kingdom was included in volume III (*Podział ludności…*, 1904). A dozen or so smaller supplementary compilations were also published, including statistics on the settlements of workers and servants, divided into groups according to occupation and place of birth.[13]

5. Assessment of the credibility of the collected data

When assessing the value of the census material in the 1920s, Szulc (1920, p. 19) wrote

> The usual reservations regarding the Russian census are legitimate: the census was conducted with the use of bureaucratic methods, the project was developed completely without the involvement of the representatives of the society, and the engagement of the public to ensure the success of the census was almost fiction. The work was often done carelessly and without due diligence. The very assumption was erroneous, there was no deeper plan in the final preparation. […] The same considerations also showed that it was an undertaking on a European scale, which nevertheless met the basic requirements of the science of demography. It was the subjective factors that were more likely to have been failing, while objective factors, so those which should minimise the impact of subjective errors, were generally standardised, though correctly, but not perfectly.

[13] The available census cards (Tsentral'nyy Statisticheskiy Komitet, 1904) containing data related to the surveyed individuals prove invaluable for demographic and, above all, genealogy research. Unfortunately, according to the findings of the central authorities, all the census cards collected in St. Petersburg were destroyed, as was the archival material collected in the governorate cities. Ultimately, however, the source material relating to some parts of the Empire was preserved. Unfortunately, there is no study that would provide comprehensive information about the existing resources. What is known is that census cards are preserved in e.g. the National Archives in Vilnius, Riga and Grodno. With regard to the lands of the 19th century Kingdom of Poland, the author of this text confirmed the presence of census cards for the Łomża District (see Figure). Digital copies of the cards from various governorates of the former Russian Empire can also be found on the website of a genealogy organisation managed by the Church of Jesus Christ of Latter-day Saints (users must first log in to access the resources collected there): https://www.familysearch.org/pl/.

According to experts, among the three objective factors which determine the value of a population census, i.e. (1) the efficiency and good will of the governing and executive bodies, (2) the population's mental level and attitude to the census, and (3) the method of processing the material, the first factor proved to be particularly neglected. As regards the Kingdom of Poland, a negative role in this respect was undoubtedly played by the Warsaw Statistical Committee, headed by Professor Simonenko. His actions lacked objectivism and were dominated by a pro-Russian attitude 'for the good of the fatherland' in terms of propaganda, which also attracted greater interest and involvement of the central administration authorities (e.g. General-Governor Aleksander Imeretyński). During the very census of 1897, certain errors were reported (documented in the report of the aforementioned census observer, Baracz): the exclusion of children, counting some people twice, reporting men who did not live with their wives as widowers, reporting age incorrectly, and others. Among the general mistakes made during the census inaccuracies in age reporting were particularly noticeable. These included age heaping, i.e. the respondents' age was rounded, especially to full decades[14] or the age provided by the elderly was higher than it actually was. The verification of the data was conducted on the basis of additional documentation which demonstrated that there were five times fewer people over the age of 100 than the census showed (Szulc, 1920).

Women more often than men tended to be excluded from the census. The wives' personal details were not always carefully noted and often the husbands declared what follows: 'Budu â veličat' ee! Baba tak i est', i net ej bol'še nazvaniâ' (I will not even name her! A woman is just a woman and there is no better name for her; Anan'eva, n.d.). Religious denomination, on the other hand, was generally reported correctly. An exception in this respect noted in the Kingdom of Poland was the census of the Uniate population in the Siedlce Governorate. As a rule, the native language was reported correctly, apart from the Jewish faith, whose followers were almost always classified as Yiddish. Occupation statistics were reported inaccurately, which was caused by the imprecision of the questionnaire (e.g. category – job position, secondary occupation; Szulc, 1920).

The actions of the authorities in the capital of the Kingdom of Poland seemed to have been consistent with the aforementioned pro-Russian policy. Their aim was to demonstrate the substantial share of Russians in the population residing in the area, which was confirmed by the denominational composition and the distribution by native language. For this purpose, barracked army was counted as part of the city's population and the area of Warsaw was expanded to include suburban towns with a large number of barracked troops. The composition of the population by native language was also distorted, as declaring Polish as their mother tongue by Evangelicals and Jews was undermined (the scale of these errors is difficult to estimate). As mentioned above, the age statistics can be considered faulty to some

[14] For more details on inaccuracies of the source data related to age heaping in historical populations, please see: Szołtysek et al. (2018).

extent, which resulted from the construction of the question requiring information on the number of years lived, and not the date or year of one's birth. These distortions most often consisted in age heaping to years ending with five or full tens. Mistakes were also found in questions relating to job positions. Attention is also drawn to the insufficient share of workers and servants in the total population, which resulted from a number of factors, including mistaking occupation for a job position and the reluctance to admit to holding positions considered the lowest in the social hierarchy (Nietyksza, 1971).

6. Conclusions

The aim of the article was to discuss the 1897 Census in the Kingdom of Poland. Its origins, the course of data collection and the processing of the results were presented, also with reference to decisions made at the central level.

The idea to prepare a population census which would cover the area of the entire Empire, i.e. approximately 22 million square kilometres, was discussed as early as in the 1860s and 1870s. The final plans were implemented in 1897. In the case of the Kingdom of Poland, formally one of the provinces of Russia, the authorities of individual governorates and the Warsaw Statistical Committee played an active role in the preparation of the census. The burden of obtaining the data rested with the lowest-level administrative authorities, which bore a number of responsibilities, including the recruitment of census enumerators. The basic census questionnaire consisted of 14 questions. The main reservations as to the credibility of the data obtained in the census concern the held position, age reporting and in the case of the data for the Kingdom of Poland the faulty results relating to religious denomination and native language.

When using the demographic data of the 1897 Census, one should bear in mind the aforementioned flaws, or even errors, which occurred both at the stage of data collecting and processing, and in the course of their publication. However, assuming a critical approach in the research of these source materials allows for an effective use of the wealth of information collected at that time.

References

Anan'eva, O. (n.d.). *Pervaâ vseobŝaâ perepis' v Rossii*. Retrieved April 10, 2021, from http://informat444.narod.ru/museum/pres/pl-6-99.htm.

Bailar, B. A. (2001). Censuses: Comparative International Aspects. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences Reference Work* (pp. 1595–1599). Elsevier. https://doi.org/10.1016/B0-08-043076-7/00398-3.

Bardach, J., Leśnodorski, B., & Pietrzak, M. (1994). *Historia ustroju i prawa polskiego*. Wydawnictwo Naukowe PWN.

Bashe, C. J., Johnson, L. R., Palmer, J. H., & Pugh, E. W. (1985). *IBM's Early Computers*. MIT Press.

Bazylow, L. (1970). *Dzieje Rosji 1801–1917*. Państwowe Wydawnictwo Naukowe.

Berger, J. (2002). Spisy ludności na ziemiach polskich do 1918 r. *Wiadomości Statystyczne*, *47*(1), 12–19.

Berger, J. (2017). Statystyka w Królestwie Polskim. In F. Kubiczek (Ed.), *Historia Polski w liczbach. Statystyka Polski. Dawniej i dzisiaj* (pp. 85–108). Główny Urząd Statystyczny. https://stat.gov.pl/files/gfx/portalinformacyjny/pl/defaultaktualnosci/5501/34/1/1/historia_polski_w_liczbach_tom_czwarty_statystyka_polski.pdf.

Bornstein, B. (1920). *Przyczynki do statystyki Królestwa Polskiego. Analiza krytyczna danych statystycznych dotyczących ruchu naturalnego ludności byłego Królestwa Polskiego*. Główny Urząd Statystyczny.

Brûhanova, E. A., & Ivanova, N. P. (2019). Dokumentovedčeskij i istočnikovedčeskij analiz pervičnyh materialov Pervoj vseobŝej perepisi naseleniâ Rossijskoj imperii 1897 g. *Vestnik Tomskogo Gosudarstvennogo Universiteta*, (442), 96–107. https://doi.org/10.17223/15617793/442/12.

Census in Rural Russia. (1897, July 12). *New York Times*, (5).

Clem, R. S. (2016). On the Use of Russian and Soviet Censuses for Research. In R. S Clem (Ed.), *Research Guide to the Russian and Soviet Censuses*, (pp. 17–35). Cornell University Press. https://doi.org/10.7591/9781501707087-004.

Craumer, P. R. (1986). Index and Guide to the Russian and Soviet Censuses, 1897 to 1979. In R. S. Clem (Ed.), *Research Guide to the Russian and Soviet Censuses* (pp. 173–323). Cornell University Press.

Domschke, E. M., & Goyer, D. S. (1986). *The Handbook of National Population Censuses. Africa and Asia*. Greenwood Press.

Edlund, T. K. (1999). The 1st National Census of the Russian Empire. *FEEFHS Journal*, *7*(3–4), 88–97. https://feefhs.org/sites/default/files/feefhs_journals/vol_07_3-4_1999.pdf.

Główny Urząd Statystyczny. (2002). *213 lat spisów ludności w Polsce 1789–2002*. https://100latgus.stat.gov.pl/historia-gus/spisy-ludnosci.

Goyer, D. S., & Domschke, E. M. (1983). *The Handbook of National Population Censuses. Latin America and the Caribbean, North America, and Oceania.* Greenwood Press.

Goyer, D. S., & Draaijer, G. E. (1992). *The Handbook of National Population Censuses. Europe*. Greenwood Press.

Grossman, H. (1925). *Struktura społeczna i gospodarcza Księstwa Warszawskiego na podstawie spisów ludności 1808–1810 r.* Główny Urząd Statystyczny.

Holzer, J. Z. (1999). *Demografia*. Polskie Wydawnictwo Ekonomiczne.

Janczak, J. K. (1982). *Ludność Łodzi przemysłowej 1820–1914*. Uniwersytet Łódzki.

Kreyenfeld, M., Willekens, F. (2015). Data Bases and Statistical Systems: Demography. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (2nd edition) (pp. 735–741). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.41013-5.

Krnilov, A. A. (1905). *Krest'ânskaâ Reforma*.

Krzyżanowski, A., & Kumaniecki, K. (1915). *Statystyka Polski*. Polskie Towarzystwo Statystyczne.

Litvak, B. G. (1991). *Perevorot 1861 goda v Rossii: počemu ne realizovalas' reformatorskaâ al'ternativa*. Politizdat.

Łukasiewicz, J. (2009). Spisy ludności w Polsce i na ziemiach polskich do 1939 r. *Wiadomości Statystyczne*, *54*(6), 1–5.

Młodak, A. (2011). *Zarys dziejów obserwacji statystycznych w Południowej Wielkopolsce*. Wydawnictwo Uczelni Państwowej Wyższej Szkoły Zawodowej im. Prezydenta Stanisława Wojciechowskiego.

Nietyksza, M. (1971). *Ludność Warszawy na przełomie XIX i XX wieku.* Państwowe Wydawnictwo Naukowe.

Podział ludności według rodzajów głównych zajęć i grup wieku w poszczególnych okręgach terytorialnych. (1904). Vol. 1–4.

Pruss, W. (1977). Skład wyznaniowo-narodowościowy ludności Warszawy w XIX i początkach XX w. In J. Kazimierski (Ed.), *Społeczeństwo Warszawy w rozwoju historycznym* (pp. 372–388). Państwowe Wydawnictwo Naukowe.

Robbins, R. G. (1975). *Famine in Russia, 1891–92. The Imperial Government Responds to a Crisis.* Columbia University Press.

Schwartz, L. (1986). A History of Russian and Soviet Censuses. In R. S. Clem (Ed.), *Research Guide to the Russian and Soviet Censuses* (pp. 48–69). Cornell University Press.

Suder, W. (2003). *Census populi. Demografia starożytnego Rzymu*. Wydawnictwo Uniwersytetu Wrocławskiego.

Szołtysek, M., Poniat, R., & Gruber, S. (2018). Age heaping patterns in Mosaic data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *51*(1), 13–38. https://doi.org/10.1080/01615440.2017.1393359.

Szulc, S. (1920). *Wartość materiałów statystycznych dotyczących stanu ludności b. Królestwa Polskiego*. Główny Urząd Statystyczny.

Thorvaldsen, G. (2018). *Censuses and census takers. A Global History*. Routledge Taylor & Francis Group.

Tsentral'nyy Statisticheskiy Komitet. (1904). *Pervaa vseobŝaa perepis' naselenia Rossijskoj imperi 1897 goda*. Vol. 1–89.

Zestawienie ogólne rezultatów spisu. (1905). Vol. 1–2.

Wiadomości Statystyczne. The Polish Statistician, 2022, vol. 67, 2, 51–61
Dyskusje. Recenzje. Informacje / Discussions. Reviews. Information

Joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality

W dniach 1–3 grudnia 2021 r. na Uniwersytecie Ekonomicznym w Poznaniu (UEP) odbyła się międzynarodowa konferencja naukowa *Joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality*, poświęcona poufności danych statystycznych. Została zorganizowana przez Eurostat oraz Europejską Komisję Gospodarczą (UNECE), a z polskiej strony przez Katedrę Statystyki UEP we współpracy z Głównym Urzędem Statystycznym (GUS) oraz Urzędem Statystycznym (US) w Poznaniu. To cykliczne wydarzenie organizują różne ośrodki na świecie, a tym razem na miejsce spotkania wybrano Poznań. Konferencję otworzyli rektor UEP Maciej Żukowski i prezes GUS Dominik Rozkrut. W tematykę obrad wprowadził Taeke Gjaltema, kierujący zespołem ds. zarządzania i modernizacji statystyki w Departamencie Statystyki Europejskiej Komisji Gospodarczej.

Program konferencji obejmował sześć sesji tematycznych zogniskowanych wokół najważniejszych wyzwań dotyczących poufności danych statystycznych, przed którymi stoi obecnie statystyka publiczna na całym świecie, tj.:

- *Access to microdata* (*Dostęp do mikrodanych*);
- *Risk assessment: Privacy, confidentiality, and disclosure* (*Ocena ryzyka: prywatność, poufność i ujawnianie*);
- *Software tools for statistical data confidentiality* (*Oprogramowanie stosowane w celu zapewnienia poufności danych statystycznych*);
- *Microdata protection* (*Ochrona mikrodanych*);
- *Tabular data* (*Dane tabelaryczne*);
- *Other emerging issues* (*Inne pojawiające się problemy*).

Łącznie wygłoszono 33 referaty.

Pierwsza sesja tematyczna *Access to microdata*, obejmująca siedem referatów, została zorganizowana przez Aleksandrę Bujnowską (Eurostat) i Erica Schulte Nordholta (Statistics Netherlands, Holandia), który był również autorem referatu inaugurującego obrady. Wystąpienie *Access to different kinds of Statistics Netherlands' microdata* poświęcono procesowi publikacji danych statystycznych i podkreślono w nim szczególny wymiar poufności mikrodanych uwzględniany w holenderskiej statystyce publicznej.

W referacie *Creating ready-made research datasets from national administrative registers* Päivi Kankaanranta i Aino Melakari (Statistics Finland, Finlandia) przedstawili projekt udostępniania danych w formie modułów zawierających standardowy zestaw zmiennych, z informacjami na poziomie firmy, gospodarstwa domowego lub osoby. Dokonali szczegółowej charakterystyki właściwości tworzonych modułów

i zauważyli, że w praktyce można z nich korzystać wyłącznie za pośrednictwem bezpiecznego systemu zdalnego dostępu, a każdy plik wyjściowy jest sprawdzany pod kątem poufności danych.

Wystąpienie Tanji Šarčević i Rudolfa Mayera (SBA Research, Austria) oraz Andreasa Raubera (University of Technology, Austria) zatytułowane *Fingerprinting relational data* dotyczyło kontroli dostępu do danych za pomocą odcisku palca. Autorzy przedstawili propozycję wykorzystania bazy danych zawierającej informacje o odciskach palców, uwzględniającej zmienne kategorialne. Nakreślili ponadto, jak mógłby wyglądać proces oceny jakości i użyteczności tak skonstruowanej bazy danych. W referacie wskazano na potrzebę znalezienia kompromisu pomiędzy niezawodnością a użytecznością metody kontroli dostępu do danych opartej na wykorzystaniu odcisku palca.

W prezentacji *Shedding light on the legal approach to aggregate data under the GDPR & the FFDR*, przedstawionej przez Emanuelę Poddę (University of Bologna, Włochy), poruszono kwestię konieczności określenia umocowania prawnego w odniesieniu do ryzyka związanego z wykorzystaniem danych zagregowanych. Chodziło przede wszystkim o wskazanie na rozbieżności pomiędzy sferą publiczną a prywatną w tym zakresie. Zwrócono również uwagę na istnienie szarej strefy w obszarze regulacji dotyczących procesu przetwarzania i udostępniania danych.

Roxane Silberman i Maria Alkhoury (Secure Data Hub, Francja) w referacie *Transnational access to confidential microdata* również poruszyły kwestię prawnych aspektów udostępniania danych. Prelegentki zwróciły uwagę na to, w jaki sposób CASD (Centre d'Accès Sécurisé aux Données – grupa interesu publicznego zrzeszająca kraje reprezentowane przez INSEE, GENES, CNRS, École Polytechnique i HEC Paris) wspiera międzynarodową współpracę badawczą, w szczególności na tle ostatnich zmian w obszarze ponadnarodowego dostępu do zabezpieczonych danych. W wystąpieniu przeanalizowano jakościowo różne zastosowania danych w prowadzonych projektach, włączając w to dostęp do danych i współpracę między instytucjami z wielu krajów. Podkreślono potrzebę dalszego rozwoju w tym zakresie, zwłaszcza z punktu widzenia użytkowników danych.

Elizabeth Green i Felix Ritchie (University of the West of England, Wielka Brytania) przygotowali wystąpienie *Microdata access*: *where we are and where we need to go* prezentujące wnioski sformułowane podczas warsztatu zorganizowanego m.in. przez University of the West of England, Europejską Komisję Gospodarczą, Eurostat i INXED-ę. Warsztat poświęcony był doświadczeniom zdobytym w ciągu ostatnich 15 lat w zakresie udostępniania mikrodanych. W referacie zaproponowano rozwiązania mające na celu przezwyciężanie praktycznych trudności w definiowaniu strategii i systemów dostępu do danych.

Natalia Volkow (National Institute of Statistics and Geography, Meksyk) w wystąpieniu *Microdata access services coping with COVID-19 lockdown* udzieliła odpowiedzi na pytanie, jakie działania zostały podjęte, aby umożliwić badaczom dostęp do mikrodanych podczas lockdownu spowodowanego przez pandemię COVID-19. W referacie uwypuklono lokalny kontekst udostępniania danych, charakteryzujący się szczególnymi cechami, które powinny być uwzględniane w celu zapewnienia zgodnego z prawem dostępu do mikrodanych. Chodzi przede wszystkim o wykorzystanie informacji statystycznych wyłącznie do celów badawczych oraz zdefiniowania, prowadzenia i oceny prowadzonych polityk publicznych.

Organizatorami drugiej sesji tematycznej *Risk assessment: Privacy, confidentiality, and disclosure* byli Josep Domingo-Ferrer (University of Rovira i Virgili, Hiszpania) oraz Krishnamurty Muralidhar (University of Oklahoma, Stany Zjednoczone).

Susan Krueger, Esma Mansouri-Benssassi (University of Dundee, Wielka Brytania) oraz Felix Ritchie i Jim Smith (University of the West of England) przygotowali inaugurujący tę sesję referat *Statistical disclosure control for machine learning models*. W prezentacji wykazano, że zarówno w sferze teoretycznej, jak i praktycznej istnieje ryzyko związane z udostępnianiem specyfikacji modelu sztucznej inteligencji z kontrolowanego środowiska. Wskazano również na rolę kontroli statystycznych (technicznych), mających na celu zmniejszenie ryzyka, oraz kontroli operacyjnych, które mogą okazać się istotne dla środowisk o określonej specyfice. Podkreślono, że wciąż mamy do czynienia z bardzo dużym stopniem niepewności, co dokładnie oznacza termin *ujawnianie* w modelach uczenia maszynowego.

W wystąpieniu *Disclosure metrics born from statistical evaluations of data utility*, którego autorami byli Devyani Biswal (University of Ottawa, Kanada), Luk Arbuckle (Privacy Analytics, Kanada) i Rafal Kulik (University of Ottawa), przedstawiono wyniki badań dotyczących wpływu różnych rozkładów szumu na statystyczne właściwości zbiorów mikrodanych poddanych tym zakłóceniom. Dzięki wykorzystaniu statystycznych testów zgodności i wybranych miar ryzyka zaproponowano nową technikę randomizacji, która poprawia użyteczność danych i jednocześnie zapewnia porównywalny poziom ryzyka ujawnienia.

Ryzyko naruszenia poufności danych spisowych poprzez zastosowanie techniki rekonstrukcji bazy danych było tematem referatu *Risk assessment procedures for the 2020 U.S. census* autorstwa Stevena Rugglesa, Lary Cleveland oraz Davida Van Ripera (Institute for Social Research and Data Innovation, University of Minnesota, Stany Zjednoczone). Okazało się, że dzięki wykorzystaniu dziewięciu tablic ze spisu z 2010 r. można odtworzyć hipotetyczną bazę danych jednostkowych z trafnością 46%. Dlatego dla rundy spisów 2020 amerykańskie biuro spisowe zaplanowało zastosowanie nowego podejścia, zakładającego dodanie szumu w przypadku każdej

charakterystyki publikowanej dla przekrojów terytorialnych na poziomie niższym niż stan.

W wystąpieniu Jesúsa Gonzáleza i Luisa Aréchigi (National Institute of Statistics and Geography, Meksyk) zatytułowanym *Proposal for a risk assessment scale for privacy risks in the disclosure of statistical information* omówiona została skala stosowana w meksykańskim krajowym urzędzie statystycznym, która służy do jakościowej oceny i analizy ryzyka bezpośredniej lub pośredniej identyfikacji jednostek w opublikowanych przez urząd danych statystycznych.

Ostatni referat w tej sesji tematycznej, wygłoszony przez Josepa Dominga-Ferrerę oraz Krishnamurty'ego Muralidhara, pt. *Database reconstruction is very difficult in practice* stanowi ważny głos w dyskusji na temat przydatności różnych metod zapewnienia poufności danych. Okazuje się, że w niektórych sytuacjach proste metody SDC mogą być bardzo skuteczne, a co więcej, mechanizmu różnicowania prywatności nie należy traktować jako panaceum na wszystkie problemy poufności danych. Ważna w tym kontekście jest uwaga Muralidhara, która padła podczas dyskusji po przedstawieniu referatu: „właściwe zastosowanie metod SDC to sztuka, a nie gotowa recepta".

Drugi dzień obrad rozpoczął się od sesji *Software tools for statistical data confidentiality*. Obejmowała ona kilka wystąpień prezentujących różnorodne rodzaje oprogramowania przeznaczonego do przeprowadzania kontroli ujawniania danych – od narzędzi do weryfikacji danych wynikowych (ang. *output checking*) przez pakiety R po język programowania wspomagający obliczenia wielostronne. Współorganizatorami sesji byli Peter-Paul de Wolf (Statistics Netherlands) i Andrzej Młodak (US w Poznaniu, Polska).

Wystąpienie Daniela P. Luppa i Øyvinda Langsruda (Statistics Norway, Norwegia) pt. *Suppression of directly-disclosive cells in frequency tables* dotyczyło postępowania w zakresie ukrywania w tablicach częstości tych komórek, które stwarzają bezpośrednie ryzyko identyfikacji jednostki. Zaproponowano nową metodę ukrywania komórek, nazwaną ukrywaniem gaussowskim. Zasygnalizowano również możliwość rozszerzenia omawianych metod na tablice hierarchiczne lub nawet na dość często spotykane tablice z ukrytymi strukturami hierarchicznymi. Nadmieniono, że rozwiązania zaczęto już wdrażać.

Michel Reiffert, Sarah Giessing i Sven Grunwald (Destatis, Niemcy) w referacie *Introducing a graphical user interface for creating the metadata governing the secondary cell suppression process* zajęli się graficznym interfejsem użytkownika (ang. *graphical user interface* – GUI) wykorzystywanym przez Federalny Urząd Statystyczny Niemiec do usprawnienia działań w zakresie spójnego ukrywania danych dla ogólnie powiązanych tablic. Część tego GUI – napisanego w języku Java i łączącego

algorytmy SAS z τ-Argus – jest używana przez specjalistów tworzących owe tablice, a inną część wykorzystują eksperci z zakresu SDC.

Gillian Raab, Beata Nowok i Chris Dibben (Scottish Centre for Administrative Data Research, Wielka Brytania) w referacie *Assessing, visualizing and improving the utility of synthetic data* omówili ostatnie ulepszenia pakietu synthpop środowiska R, służącego do tworzenia danych syntetycznych. Dotyczą one zwłaszcza miar użyteczności wygenerowanych zbiorów, okazało się bowiem, że pewne umieszczone tam mierniki mierzą w zasadzie to samo. Autorzy przedstawili również działania dotyczące oceny użyteczności danych z wykorzystaniem technik wizualizacyjnych, a nie tylko bazujących na liczbach.

Giuseppe Bruno (Bank of Italy, Włochy) w wystąpieniu *Private linear regression: Can we scale up with Big Data?* rozważał możliwości zastosowania języka programowania „obliczeń nieświadomych", czyli działań analitycznych (np. analizy regresji na danych indywidualnych ze zbiorów big data) z wykorzystaniem bibliotek *open source*. Zaprezentował, jak przeprowadzać wielostronne obliczenia z użyciem języka Obliv-C, mimo że jego rozwój nie jest jeszcze aktywnie wspierany.

Na zakończenie tej sesji Marco Stocchi (Eurostat) i Aleksandra Bujnowska w referacie *Automatic checking of research outputs* przedstawili pilotażową wersję oprogramowania ACRO służącego do przeprowadzania kontroli danych wynikowych. Jest ono przeznaczone dla użytkowników środowiska STATA i ma na celu wykrywanie problemów z ochroną tajemnicy statystycznej, a nie ich rozwiązywanie.

W trakcie dyskusji dotyczącej powyższych wystąpień zastanawiano się m.in. nad tym, czy graficzny interfejs umożliwi przeprowadzanie procesu kontroli ujawniania danych dla tablic z różnych badań oraz czy w przypadku danych syntetycznych zaproponowany model będzie efektywny dla tablic wielowymiarowych. Pytano również o powszechność użycia niektórych środowisk, w których opracowano określone algorytmy. Wzmianka o SAS i STATA wywołała również krótką debatę na temat cech odróżniających oprogramowanie typu *open source* od oprogramowania komercyjnego.

Czwarta sesja została poświęcona ochronie mikrodanych. Jej organizatorami byli Josep Domingo-Ferrer i Krishnamurty Muralidhar. Składało się na nią sześć referatów. Większość dotyczyła danych syntetycznych lub wykorzystania sztucznej inteligencji.

Sana Rashid (University of Southampton, Wielka Brytania), Jörg Drechsler (Institute for Employment Research, Niemcy) i Robin Mitra (University of Cardiff, Wielka Brytania) w referacie *Accounting for longitudinal data structures when disseminating synthetic data to the public* poruszyli problematykę koncepcji generowania

syntetycznych danych wzdłużnych i ocenili efektywność kilku służących temu strategii.

Johannes Gussenbauer i Alexander Kowarik (Statistics Austria, Austria) oraz Klaudius Kalcher i Michael Platzer (Mostly AI, Austria) w referacie *AI-based privacy preserving census(like) data publication* omówili podejście polegające na wykorzystaniu sztucznej inteligencji (głębokich generatywnych sieci neuronowych) do tworzenia syntetycznych danych spisowych z zapewnieniem ochrony tajemnicy statystycznej.

Generating tabular data using generative adversarial networks with differential privacy było tematem wystąpienia Giacoma Astolfiego i Davida Küpfera (European Central Bank, Niemcy). Autorzy przedstawili metodę tworzenia syntetycznych danych tabelarycznych chroniących prywatność przy użyciu zróżnicowanych generatywnych sieci kontradyktoryjnych (ang. *generative adversarial neural nets* – GAN) oraz połączenia tej metody z różnicowaniem prywatności (ang. *differential privacy* – DP).

Claire Little, Mark Elliot, Richard Allmendinger i Sahel Shariati Samani (University of Manchester, Wielka Brytania) w referacie *Generative adversarial networks for synthetic data generation: A comparative study* omówili użycie sieci GAN do generowania danych syntetycznych i porównali to podejście z innymi metodami, takimi jak te zawarte w pakiecie synthpop środowiska R.

Kenza Sallier (Statistics Canada, Kanada) w referacie *Data access modernization in National Statistical Offices through synthetic data, the HLG-MOS guide* przedstawiła podejście do dostępu do informacji statystycznych oparte na danych syntetycznych. Zaprezentowała przy tym przewodnik dotyczący tworzenia danych syntetycznych przeznaczony dla krajowych urzędów statystycznych, który powstał w ramach projektu UNECE HLG-MOS, będącego pokłosiem dyskusji podczas spotkania przedstawicieli UNECE i Eurostatu w sprawie poufności danych statystycznych, które odbyło się w Hadze w 2019 r.

Oryginalne podejście do syntezy wartości ekstremalnych w złożonych zbiorach danych przedstawili Anna Oganian (National Center for Health Statistics, Stany Zjednoczone), Mehtab Iqbal (Clemson University, Stany Zjednoczone) i Goran Lesaja (Georgia Southern University, United States Naval Academy, Stany Zjednoczone) w referacie *Extreme value protection adjustment for different subpopulations in complex data sets*.

Dyskusja tocząca się w tej sesji niemal w połowie dotyczyła zagadnienia uczenia maszynowego (w tym mechanizmu czarnej skrzynki) oraz jego zastosowań statystycznych. Podobnie jak w przypadku analizy regresji, gdy niektórzy użytkownicy są zainteresowani jedynie końcowymi predykcjami (i tu czarna skrzynka jest właściwa),

a inni – wnioskowaniem (w przypadku którego konieczne jest podejście statystyczne), również użytkownicy danych syntetycznych mają zróżnicowane potrzeby. Tym samym podkreślono istotność rozważenia i rozpoznania kierunku zastosowania danych syntetycznych oraz sposobu działania tych dwóch podejść (tzn. uczenia maszynowego i statystycznego) w tym kontekście. Dyskutanci poruszali także kwestię różnicy między danymi panelowymi a klasycznymi danymi wielowymiarowymi z punktu widzenia kontroli ujawniania danych oraz uwzględnienia typowych metod tej kontroli w sieciach kontradyktoryjnych.

Wystąpienia wygłoszone podczas sesji *Tabular data*, której organizatorami byli Steven Thomas (Statistics Canada) i Sarah Giessing, poświęconej ochronie danych tabelarycznych zapewniły doskonały przegląd wyzwań napotykanych podczas publikowania informacji tabelarycznych przez instytucje statystyczne. Każda prezentacja zawierała odniesienie do kluczowego w tym kontekście balansu pomiędzy ryzykiem ujawnienia a użytecznością udostępnianych informacji. Ryzyko może być realne lub postrzegane, często trudne do zmierzenia, a użyteczność zależy od punktu widzenia użytkownika, zatem osiągnięcie równowagi pomiędzy ryzykiem a użytecznością jest trudnym zadaniem.

Sesję otworzył referat Fabiana Bacha (Eurostat) pt. *Differential privacy and noisy confidentiality concepts for European population statistics*, w którym autor omówił metodę kluczy komórkowych (ang. *cell-key method*) i jej kontrast z mechanizmem różnicowania prywatności DP. Użycie parametru epsilon (zakładanej redukcji ryzyka ujawnienia) może być wykorzystywane do osiągnięcia równowagi z użytecznością informacji. Ryzyko można określić ilościowo, zwłaszcza biorąc pod uwagę masowe uśrednianie danych. Wydaje się, że nałożenie szumu wygenerowanego na podstawie podejścia DP czyni szacunki bezużytecznymi ze względu na problem nieograniczoności szumu. Ostatecznie do ochrony danych zgromadzonych w spisach powszechnych 2021 r. zaleca się metodę kluczy komórkowych. Prelegent podkreślił, że nie ma możliwości dokonania prostego wyboru parametrów, który mógłby zapewnić jednocześnie akceptowalny poziom ryzyka ujawnienia i użyteczności wyników statystycznych.

Øyvind Langsrud i Daniel P. Lupp w referacie *Fair risk-utility comparison of tabular perturbation methods by post-processing to expected frequencies* powrócili do zagadnienia równowagi między użytecznością a ryzykiem, analizując pod tym względem metodę kluczy komórkowych oraz ideę strategii opartej na zaokrąglaniu małych liczb. W wystąpieniu ukazano miary ryzyka, które można wykorzystać do porównywania wyników uzyskiwanych po zastosowaniu obu podejść pod kątem ochrony tajemnicy statystycznej. Zwrócono też uwagę na wyzwania związane z określeniem

akceptowalnych progów ryzyka oraz na to, że nie ma jednej miary ryzyka, którą można by uznać za najlepsze rozwiązanie.

Pytanie zawarte w tytule referatu Wilhelmusa Kloeka (Eurostat) *Suppression or perturbation?* dotyczyło dylematu, czy komplementarna strategia ukrywania komórek jest odpowiednia, czy też lepszym rozwiązaniem byłoby zakłócanie danych wedle metody kluczy komórkowych. Autor omówił zalety i wady tych metod oraz zauważył, jak ważne jest uwzględnienie błędów losowych i nielosowych podczas stosowania tych metod, co wyraźnie łączy się ze znaczeniem przejrzystości, gdy nakłada się zakłócenia.

Sarah Giessing, Michel Reiffert i Felix Geyer (Destatis) oraz Peter-Paul de Wolf w referacie *Considerations to deal with the frozen cell problem in Tau-Argus Modular* przedstawili wyzwania związane z opracowywaniem spójnych wzorców ukrywania komórek w sytuacji, gdy tabele, których opracowania oczekują użytkownicy, mają zostać opublikowane po upublicznieniu tabel, w których już użyto komplementarnej strategii ukrywania komórek w celu zachowania tajemnicy statystycznej. W tych nowych tabelach komórki poprzednio ukryte muszą pozostać ukryte, a komórki już opublikowane nie mogą być ukrywane wtórnie. Reprezentująca autorów Giessing omówiła sposób ulepszenia podejścia modułowego wdrożonego w programie τ-Argus, aby umożliwić realizację niewykonalnych dotąd działań. Zwróciła również uwagę na wyzwanie polegające na osiągnięciu równowagi ryzyka i użyteczności, w której można zaakceptować pewne słabsze zabezpieczenia dla udostępnienia użytecznych, ale konsekwentnie chronionych i połączonych tabel.

Kwestie związane z udostępnianiem zagregowanych statystyk ekonomicznych były przedmiotem referatu *Increasing utility of economic statistical information* Stevena Thomasa (Statistics Canada). Autor wykazał ponadto, że strategia ukrywania komórek, przedstawiona przez Giessing, powoduje pewne problemy dotyczące użyteczności, a zakłócanie przedstawił jako alternatywę. Zaletą tego drugiego podejścia jest możliwość sprawnego opracowywania miar ryzyka i użyteczności. Z tego względu – jak podkreślił – w kanadyjskiej statystyce gospodarczej obserwuje się tendencję do odchodzenia od ukrywania wrażliwych danych na rzecz ich zakłócania.

Podczas dyskusji na temat powyższych wystąpień zauważono, że przyszłe prace dotyczące kontroli ujawniania danych powinny poruszać kwestie opracowania standardowych miar ryzyka, które można będzie wykorzystywać do pomiaru ryzyka i umożliwienia porównania różnych strategii ochrony tajemnicy statystycznej stosowanych do danych zagregowanych. Może to wymagać współpracy międzynarodowej w zakresie zdefiniowania istoty ryzyka i użyteczności oraz wypracowania sposobu komunikowania tych kwestii wewnętrznym i zewnętrznym interesariuszom (użytkownikom). Sygnalizowano również, że chociaż badacze wolą otrzymywać oryginal-

ne i niezaburzone dane, to sprawdzanie wyników badań (na podstawie oryginalnych danych) może być bardziej skomplikowane niż w przypadku zakłóconych wyników publikowanych przez instytucje statystyczne. Zauważono, że w różnych krajach obowiązują odmienne początkowe założenia dotyczące poufności (np. w Stanach Zjednoczonych za potencjalnie zagrożone ujawnieniem uznaje się agregaty liczące mniej niż 10 jednostek).

W ostatnim dniu konferencji zorganizowano sesję poświęconą różnym innym aspektom i problemom kontroli ujawniania danych. Jej współorganizatorami byli Peter-Paul de Wolf i Janika Tarkoma (Statistics Finland). W sesji wygłoszono trzy referaty.

W wystąpieniu *A proof-of-concept solution for secure processing of mobile network operator data for official statistics* Fabio Ricciato i Albrecht Wirthmann (Eurostat) oraz Triin Siil, Riivo Talviste i Baldur Kubo (Cybernetica, Estonia) przedstawili główne założenia projektu realizowanego przez Eurostat i firmę Cybernetica (dostawcę rozwiązań wykorzystywanych do bezpiecznej analizy danych) w latach 2020–2021, którego podstawowym celem było zastosowanie rozwiązania umożliwiającego bezpieczne przetwarzanie (ang. *secure private computing*) danych operatora sieci komórkowej z zapewnieniem zasady zachowania prywatności. Wykorzystywana na potrzeby projektu technologia zaufanego środowiska wykonawczego (ang. *trusted execution environment* – TEE) pomaga producentom, usługodawcom i konsumentom chronić urządzenia i wrażliwe dane. Jest to środowisko, w którym opracowany i wykonywany kod i przechowywane dane, do których uzyskuje się dostęp, są fizycznie izolowane i poufnie chronione. Dzięki takiemu podejściu żaden podmiot bez integralności nie może uzyskać fizycznego dostępu do danych ani zmienić kodu czy też sposobu jego zachowania. W referacie postawiono trzy zasadnicze pytania: (1) czy rozwiązania TEE są wystarczająco skalowalne, aby poradzić sobie z dużymi wolumenami danych?; (2) czy rozwiązania TEE są na tyle dojrzałe i elastyczne, aby mogły być stosowane w statystyce publicznej?; (3) jakie są uwarunkowania prawne związane ze stosowaniem technologii TEE? Przedstawione w referacie wyniki uzyskane na bazie zbiorowości abonentów telefonii komórkowej pokazały, że rozwiązania oparte na TEE z izolacją sprzętową są wystarczająco skalowalne i zdolne do przetwarzania dużych ilości danych wejściowych oraz mogą być z powodzeniem wykorzystywane w statystyce publicznej z uwzględnieniem obowiązujących przepisów prawa.

Muhammad Aslam Jarwar i Mark Elliot (Centre for Digital Trust and Society, University of Manchester) oraz Age Chapman i Fatemeh Raji (School of Electronics and Computer Science, University of Southampton) w referacie *Modelling data environments within PROV to assist anonymisation decision-making* rozważali zagadnie-

nia związane z podejmowaniem decyzji dotyczących anonimizacji danych w kontekście ich operacjonalizacji i zarządzania ryzykiem ich wymiany między organizacjami i środowiskami społecznymi i informatycznymi. Prelegenci omówili nowe zastosowanie procesu wsparcia anonimizacji danych i sposobu ich wymiany między różnymi podmiotami z wykorzystaniem koncepcji ADF (ang. Anonymisation Decision-making Framework), która jest obecnie dobrze ugruntowanym podejściem wspomagającym przepływy danych oraz metody ich anonimizacji. Zgodnie z uwagami autorów, aby działania w tym zakresie były skuteczne, główny nacisk należy położyć na podstawowy element ADF, czyli na środowisko danych, które oprócz samych danych obejmuje procesy zarządzania nimi oraz infrastrukturę informatyczną. W wystąpieniu zidentyfikowano kluczowe właściwości takich środowisk, które można mapować za pomocą modelu W3C PROV przeznaczonego do obsługi wymiany informacji.

W zamykającym sesję referacie *Structural uniqueness in network data* Mark van der Loo, Rachel de Jong, Frank Takes, Marieke de Vries i Peter-Paul de Wolf przedstawili nową metodę pomiaru stopnia anonimowości wierzchołków w sieci na gruncie teorii grafów. Bazuje ona na porównaniu pełnej pozycji strukturalnej wierzchołków w ich bezpośrednim sąsiedztwie i uwzględnia wielkość sąsiedztwa. Prelegenci obliczyli tę miarę w dużych sieciach na przykładzie holenderskiej sieci powiązań rodzinnych, z ponad 15 mln wierzchołków. Uzyskane dotychczas wyniki obliczeniowe odnoszą się do grafów nieoznakowanych, czyli sieci, w których wierzchołki i krawędzie nie mają żadnych właściwości. Teoretyczne i obliczeniowe rezultaty pokazują, że informacje o strukturze obiektów mogą okazać się odkrywcze. Wiedza o relacjach rodzinnych do drugiego stopnia więzi włącznie, nawet bez uwzględnienia informacji o relacji rodzic – dziecko, deanonimizuje tysiące osób (wierzchołków) w sieci powiązań rodzinnych.

Po zakończeniu ostatniej sesji Dennis Ramondt (Statistics Netherlands) omówił najważniejsze kwestie związane z projektem dotyczącym zachowania prywatności danych wejściowych (*HLG project on input privacy preservation*), którego inicjatorami były urzędy statystyczne Włoch i Holandii, a w którym biorą również udział przedstawiciele Eurostatu, UNECE czy ONS. Ramondt skupił się na przedstawieniu poszczególnych pakietów roboczych wykorzystywanych do dokumentowania studiów przypadku czy stosowania metod uczenia maszynowego w kontekście ochrony prywatności danych.

Tegoroczna konferencja odbyła się w trybie hybrydowym: 20 uczestników z zagranicy przyjechało do Poznania, a ok. 170 osób wzięło udział w wydarzeniu w formie online. Uczestnicy konferencji, w tym prelegenci, łączyli się z Poznaniem z różnych stref czasowych: od Australii przez Azję (Japonia, Wietnam), Afrykę i Europę

po Stany Zjednoczone, Kanadę i Meksyk. Było to największe z dotychczas zorganizowanych spotkań z tego cyklu. W pracach komitetu organizacyjnego ze strony polskiej wzięli udział: Grażyna Dehnel, Tomasz Józefowski, Tomasz Klimanek, Andrzej Młodak, Marcin Szymkowiak i Kamil Wilak.

Wszystkie wygłoszone referaty są dostępne na stronie internetowej UNECE https://unece.org/statistics/events/SDC2021.

Grażyna Dehnel
Uniwersytet Ekonomiczny w Poznaniu, Instytut Informatyki i Ekonomii Ilościowej; Urząd Statystyczny w Poznaniu, Ośrodek Statystyki Krótkookresowej, Polska
Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics; Statistical Office in Poznań, Centre for Short-Term Statistics, Poland
ORCID: https://orcid.org/0000-0002-0072-9681
Andrzej Młodak
Akademia Kaliska im. Prezydenta Stanisława Wojciechowskiego, Międzywydziałowy Zakład Matematyki i Statystyki; Urząd Statystyczny w Poznaniu, Ośrodek Statystyki Małych Obszarów, Polska
Calisia University – Kalisz, Interfaculty Department of Mathematics and Statistics; Statistical Office in Poznań, Centre for Small Area Estimation, Poland
ORCID: https://orcid.org/0000-0002-6853-9163
Tomasz Klimanek
Uniwersytet Ekonomiczny w Poznaniu, Instytut Informatyki i Ekonomii Ilościowej, Polska
Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Poland
ORCID: https://orcid.org/0000-0001-6411-8145
Marcin Szymkowiak
Uniwersytet Ekonomiczny w Poznaniu, Instytut Informatyki i Ekonomii Ilościowej; Urząd Statystyczny w Poznaniu, Polska
Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics; Statistical Office in Poznań, Poland
ORCID: https://orcid.org/0000-0003-3432-4364

Wiadomości Statystyczne. The Polish Statistician, 2022, vol. 67, 2, 62–63
Dyskusje. Recenzje. Informacje / Discussions. Reviews. Information

WYDAWNICTWA GUS. STYCZEŃ 2022
PUBLICATIONS OF STATISTICS POLAND. JANUARY 2022

W ofercie wydawniczej Głównego Urzędu Statystycznego z ubiegłego miesiąca warto zwrócić uwagę na następującą publikację:
Among Statistics Poland's last month's publications, we would like to recommend:

Tytuł: *Ubóstwo w Polsce w latach 2019 i 2020*
Title: *Poverty in Poland in 2019 and 2020*
Język: polski (dodatkowo przedmowa, spis treści i synteza w języku angielskim)
Language: Polish (additionally preface, contents and executive summary in English)
Dodatkowe informacje: opracowanie dostępne w wersji elektronicznej
Additional information: publication available in the electronic form

Opracowanie dostarcza informacji o zasięgu i społecznym zróżnicowaniu ubóstwa i niedostatku w Polsce. Oprócz statystyk o charakterze ogólnokrajowym zawiera dane ukazujące sytuację w naszym kraju na tle innych krajów Unii Europejskiej. Przedstawione wyniki dotyczą głównie lat 2019 i 2020 (a więc odnoszą się do okresu sprzed wybuchu pandemii COVID-19 i pierwszego roku jej trwania). Pochodzą z przeprowadzanych corocznie przez GUS ankietowych badań gospodarstw domowych: badania budżetów gospodarstw domowych oraz Europejskiego Badania Dochodów i Warunków Życia Ludności (EU-SILC). Źródłem informacji dotyczących krajów członkowskich UE są dane udostępniane przez Eurostat.

W styczniu br. ukazały się ponadto:
- *Aktywność ekonomiczna ludności Polski – 3 kwartał 2021 r.*;
- „Biuletyn statystyczny" nr 12/2021;
- *Ceny robót budowlano-montażowych i obiektów budowlanych (listopad 2021 r.)*;
- *Działalność badawcza i rozwojowa w Polsce w 2020 r.*;
- *Działalność innowacyjna przedsiębiorstw w latach 2018–2020*;
- *Energia ze źródeł odnawialnych w 2020 r.*;
- *Koniunktura w przetwórstwie przemysłowym, budownictwie, handlu i usługach 2000–2022 (styczeń 2022)*;
- *Produkcja ważniejszych wyrobów przemysłowych w grudniu 2021 r.*;

- *Sytuacja demograficzna Polski do 2020 r. Zgony i umieralność*;
- *Sytuacja społeczno-gospodarcza kraju w 2021 r.*;
- „Wiadomości Statystyczne. The Polish Statistician" nr 1/2022.

Wszystkie publikacje GUS w wersji elektronicznej są dostępne na stronie stat.gov.pl/publikacje/publikacje-a-z.

Electronic forms of all the publications by Statistics Poland are available at stat.gov.pl/en/publications.

Justyna Gustyn
Główny Urząd Statystyczny, Departament Opracowań Statystycznych
Statistics Poland, Statistical Products Department

DLA AUTORÓW
FOR THE AUTHORS

(for the English translation of the information given below, please visit ws.stat.gov.pl/ForAuthors)

W „Wiadomościach Statystycznych. The Polish Statistician" („WS") zamieszczane są artykuły o charakterze naukowym poświęcone teorii i praktyce statystycznej, które prezentują wyniki oryginalnych badań teoretycznych lub analitycznych wykorzystujących metody statystyki matematycznej, opisowej bądź ekonometrii. Ukazują się również artykuły przeglądowe, recenzje publikacji naukowych oraz inne opracowania informacyjne. W czasopiśmie publikowane są prace w języku polskim i angielskim.

Od 2007 r. „WS" znajdują się na liście polskich punktowanych czasopism naukowych MEiN. Zgodnie z komunikatem Ministra Edukacji i Nauki z dnia 1 grudnia 2021 r. w sprawie wykazu czasopism naukowych i recenzowanych materiałów z konferencji międzynarodowych „WS" otrzymały 70 punktów.

„Wiadomości Statystyczne. The Polish Statistician" są udostępniane w następujących bazach, repozytoriach, katalogach i wyszukiwarkach: Agro, BazEkon, Central and Eastern European Academic Source (CEEAS), Central and Eastern European Online Library (CEEOL), Central European Journal of Social Sciences and Humanities (CEJSH), EBSCO Discovery Service, European Reference Index for the Humanities and Social Sciences (ERIH Plus), Exlibris Primo, Google Scholar, ICI Journals Master List, ICI World of Journals, Norwegian Register for Scientific Journals and Publishers (The Nordic List) oraz Summon.

Za publikację artykułów na łamach „WS" autorzy nie otrzymują honorariów ani nie wnoszą opłat.

1. Zgłaszanie artykułów

Prace należy przesyłać na adres: **redakcja.ws@stat.gov.pl**.

Artykuł powinien być utrzymany w formie bezosobowej i zawierać streszczenie, słowa kluczowe, kod/kody JEL oraz ORCID, adres e-mail i afiliację autora. Tytuł, streszczenie, słowa kluczowe i afiliacja powinny być podane w języku polskim i angielskim.

Jeżeli w pracy występują tablice, wykresy lub mapy, powinny być umieszczone w treści artykułu. Ponadto należy je przesłać osobno (najlepiej w formacie Excel); szczegółowe informacje są zawarte w pkt 4.2.14.

Autor jest zobowiązany do podania w artykule wszelkich źródeł finansowania badań będących podstawą pracy. Jeżeli doszło do zaprezentowania podobnych materiałów podczas konferencji lub sympozjum naukowego albo jeżeli artykuł został przez autora umieszczony w repozytorium internetowym lub zgłoszony w innym wydawnictwie do opublikowania w innej wersji językowej, to podczas składania tekstu do publikacji w „WS" należy poinformować o tym redakcję.

Prosimy o niestosowanie stylów i ograniczenie formatowania do wymogów redakcyjnych. Więcej informacji w rozdziale *Wymogi redakcyjne*.

Razem z artykułem należy przesłać skan oświadczenia (do pobrania ze strony internetowej czasopisma) o oryginalności pracy, niezłożeniu jej w innym wydawnictwie i udzieleniu wydawcy „WS" licencji Creative Commons Uznanie autorstwa – Na tych samych warunkach 4.0 (CC BY-SA 4.0), zawierającego numer ORCID, afiliację lub afiliacje oraz dane kontaktowe autora, wraz ze wskazaniem proponowanego działu czasopisma.

Załączenie skanu oświadczenia jest warunkiem poddania pracy ocenie wstępnej i skierowania do recenzji.

2. Przebieg prac redakcyjnych

Zgłoszony artykuł jest oceniany i opracowywany w czteroetapowym procesie:

1. **Ocena wstępna**, dokonywana przez redakcję. Polega na weryfikacji naukowego charakteru artykułu oraz jego struktury i zawartości pod kątem wymogów redakcyjnych, a także zgodności tematyki z profilem czasopisma. Autor uzupełnia i poprawia artykuł stosownie do uwag redakcji, a w przypadku nieuwzględnienia danej uwagi uzasadnia swoje stanowisko. **Razem z poprawionym artykułem autor przesyła w osobnym pliku zanonimizowaną wersję pracy, przeznaczoną do recenzji.** Anonimizacja polega na utajnieniu nazwiska autora (także we właściwościach pliku), usunięciu podziękowań i informacji o źródłach finansowania, a także innych informacji wskazujących na afiliację lub umożliwiających zidentyfikowanie autora. Warunkiem skierowania pracy do recenzji jest potwierdzenie oryginalności tekstu uzyskane za pomocą systemu antyplagiatowego Similarity Check. W przypadku wykrycia znacznego podobieństwa do innych prac artykuł zostanie odrzucony.
2. **Ocena recenzentów**, dokonywana przez specjalistów w danej dziedzinie. Artykuł oceniają dwaj recenzenci spoza jednostki naukowej, przy której afiliowany jest autor; w przypadku pracy w języku angielskim co najmniej jeden recenzent jest afiliowany przy jednostce zagranicznej. W razie sprzecznych opinii dwóch recenzentów powoływany jest trzeci recenzent. Recenzenci kierują się kryteriami oryginalności i jakości opracowania zarówno w odniesieniu do treści, jak i formy artykułu.

   Autorzy artykułów, które otrzymały pozytywne oceny, wprowadzają poprawki zalecane przez recenzentów i przesyłają do redakcji zmodyfikowaną wersję pracy. Jeśli pojawi się różnica zdań dotycząca zasadności proponowanych zmian, autorzy są zobligowani do uzasadnienia swojego stanowiska.
3. **Ocena Kolegium Redakcyjnego** (KR), decydująca o przyjęciu pracy do publikacji. Jest dokonywana na podstawie recenzji, z uwzględnieniem opinii redaktorów tematycznego i merytorycznego. Polega m.in. na weryfikacji dokonania przez autora zmian w artykule stosownie do uwag recenzentów. KR ocenia artykuł pod względem poprawności i spójności merytorycznej oraz zaleca autorowi wprowadzenie poprawek, jeśli są one konieczne, aby praca spełniała wymogi czasopisma. Autorowi przysługuje prawo do odwołania od decyzji o niepublikowaniu artykułu. W takim przypadku powinien on skontaktować się z redakcją „WS" i przedstawić uzasadnienie. Ostateczna decyzja w tej sprawie należy do redaktora naczelnego.

   W „WS" publikowane są wyłącznie te artykuły, które otrzymają pozytywną ocenę na każdym z wymienionych etapów i zostaną poprawione przez autora zgodnie z otrzymanymi uwagami (chyba że autor przedstawi argumenty uzasadniające nieuwzględnienie danej uwagi).

   Artykuły przyjęte przez KR do publikacji są zamieszczane na stronie internetowej czasopisma w zakładce Early View, gdzie znajdują się do czasu opublikowania w konkretnym wydaniu „WS".
4. **Opracowanie redakcyjne, autoryzacja i korekta.** Artykuł zakwalifikowany do druku jest poddawany opracowaniu merytorycznemu i językowemu. Redakcja zastrzega sobie prawo

do zmiany tytułu i śródtytułów, modyfikowania tablic, wykresów i innych elementów graficznych oraz przeredagowania treści bez naruszenia zasadniczej myśli autora.

Po opracowaniu redakcyjnym artykuł jest przesyłany do autoryzacji. Tekst zatwierdzony przez autora, po składzie i łamaniu, jest poddawany korekcie i rewizji (II korekcie). Autor dokonuje korekty autorskiej tekstu na etapie rewizji. Wykresy i inne materiały graficzne są opracowywane na podstawie plików i danych przekazanych przez autora i poddawane korekcie i rewizji. Autor dokonuje ich akceptacji na etapie rewizji.

W przypadku odkrycia błędów w opublikowanym artykule zamieszcza się na łamach „WS" sprostowanie, a artykuł w wersji elektronicznej jest poprawiany i umieszczany na stronie internetowej „WS" ze stosownym wyjaśnieniem.

3. Zasady etyki publikacyjnej COPE

Redakcja „WS" podejmuje wszelkie starania w celu utrzymania najwyższych standardów etycznych zgodnie z wytycznymi Komitetu ds. Etyki Publikacyjnej (COPE), dostępnymi na stronie internetowej www.publicationethics.org, oraz wykorzystuje wszystkie możliwe środki mające na celu zapobieżenie nadużyciom i nierzetelności autorskiej. Przyjęte zasady postępowania obowiązują autorów, Radę Naukową, Kolegium Redakcyjne, redakcję, pracowników Wydziału Czasopism Naukowych GUS, recenzentów i wydawcę.

3.1. Odpowiedzialność autorów

1. Artykuły naukowe kierowane do opublikowania w „WS" powinny zawierać precyzyjny opis badanych zjawisk i stosowanych metod oraz autorskie wnioski i sugestie dotyczące rozwoju badań i analiz statystycznych. Autorzy powinni wyraźnie określić cel artykułu oraz jasno przedstawić wyniki przeprowadzonej analizy. Prezentacja efektów badań statystycznych zaprojektowanych i przeprowadzonych przez autorów wymaga opisania zastosowanej w nich metodologii. W przypadku nowatorskich metod analizy pożądane jest podanie przykładu ilustrującego ich zastosowanie w praktyce statystycznej. Autorzy ponoszą odpowiedzialność za treści prezentowane w artykułach. W razie zgłaszania przez czytelników zastrzeżeń odnoszących się do tych treści autorzy są zobligowani do udzielenia odpowiedzi za pośrednictwem redakcji.
2. Na autorach spoczywa obowiązek zapewnienia pełnej oryginalności przedłożonych prac. Redakcja nie toleruje przejawów nierzetelności naukowej autorów, takich jak:
   - duplikowanie publikacji – ponowne publikowanie własnego utworu lub jego części;
   - plagiat – przywłaszczenie cudzego utworu lub jego fragmentu bez podania informacji o źródle;
   - fabrykowanie danych – oparcie pracy naukowej na nieprawdziwych wynikach badań;
   - autorstwo widmo (*ghost authorship*) – nieujawnianie współautorów, mimo że wnieśli oni istotny wkład w powstanie artykułu;
   - autorstwo gościnne (*guest authorship*) – podawanie jako współautorów osób o znikomym udziale lub niebiorących udziału w opracowywaniu artykułu;
   - autorstwo grzecznościowe (*gift authorship*) – podawanie jako współautorów osób, których wkład jest oparty jedynie na słabym powiązaniu z badaniem.

Autorzy deklarują w stosownym oświadczeniu, że zgłaszany artykuł nie narusza praw autorskich osób trzecich, nie był dotychczas publikowany i nie został złożony w innym wydawnictwie oraz że jest ich oryginalnym dziełem, i określają swój wkład w opracowanie artykułu. Jeżeli doszło do zaprezentowania podobnych materiałów podczas konferencji lub sympozjum naukowego, to podczas składania tekstu do publikacji w „WS" autorzy są zobowiązani poinformować o tym redakcję.

3. Autorzy są zobowiązani do podania w treści artykułu wszelkich źródeł finansowania badań będących podstawą pracy.
4. Główną odpowiedzialność za rzetelność przekazanych informacji, łącznie z informacją na temat wkładu poszczególnych współautorów w powstanie artykułu, ponosi zgłaszający artykuł.
5. Autorzy zgłaszający artykuły do publikacji w „WS" biorą udział w procesie recenzji double-blind peer review, dokonywanej przez co najmniej dwóch niezależnych ekspertów z danej dziedziny. Po otrzymaniu pozytywnych recenzji autorzy wprowadzają zalecane przez recenzentów poprawki i dostarczają redakcji zaktualizowaną wersję opracowania wraz z pisemnym poświadczeniem uwzględnienia poprawek. Jeśli pojawi się różnica zdań co do zasadności proponowanych zmian, należy wyjaśnić, które poprawki zostały uwzględnione, a w przypadku ich nieuwzględnienia – uzasadnić swoje stanowisko.
6. Jeżeli autorzy odkryją w swoim maszynopisie lub tekście już opublikowanym błędy, nieścisłości bądź niewłaściwe dane, powinni niezwłocznie poinformować o tym redakcję w celu dokonania korekty, wycofania tekstu lub zamieszczenia sprostowania. W przypadku korekty artykułu już opublikowanego jego nowa wersja jest zamieszczana na stronie internetowej „WS" wraz ze stosownym wyjaśnieniem.

3.2. Odpowiedzialność Rady Naukowej, Kolegium Redakcyjnego i Wydziału Czasopism Naukowych GUS

1. Rada Naukowa (RN) kształtuje profil programowy czasopisma, określa kierunki jego rozwoju i konsultuje jego zakres merytoryczny.
2. Kolegium Redakcyjne (KR) podejmuje decyzję o publikacji danego artykułu z uwzględnieniem ocen recenzentów oraz opinii zespołu redakcyjnego. W swoich rozstrzygnięciach członkowie KR kierują się kryteriami merytorycznej oceny wartości artykułu, jego oryginalności i jasności przekazu, a także ścisłego związku z celem i zakresem tematycznym „WS". Oceniają artykuły niezależnie od płci, rasy, pochodzenia etnicznego, narodowości, religii, wyznania, światopoglądu, niepełnosprawności, wieku lub orientacji seksualnej ich autorów.
3. Zespół redakcyjny, wyodrębniony z KR, tworzą redaktor naczelny i jego zastępca, redaktorzy tematyczni i redaktor merytoryczny. Członkowie zespołu redakcyjnego weryfikują nadsyłane artykuły pod względem merytorycznym, oceniają ich zgodność z celem i zakresem tematycznym „WS" oraz sprawdzają spełnienie wymogów redakcyjnych i przestrzeganie zasad rzetelności naukowej. Ponadto wybierają recenzentów w taki sposób, aby nie wystąpił konflikt interesów, i dbają o zapewnienie uczciwego, bezstronnego i terminowego procesu recenzowania.
4. Za sprawny przebieg procesu wydawniczego, poinformowanie wszystkich jego uczestników o konieczności przestrzegania obowiązujących zasad i przygotowanie artykułów do

publikacji odpowiadają pracownicy Wydziału Czasopism Naukowych (WCN) GUS. W celu uzyskania obiektywnej oceny oryginalności nadsyłanych artykułów przed skierowaniem ich do recenzji WCN wykorzystuje system antyplagiatowy. Informacje dotyczące artykułu mogą być przekazywane przez WCN wyłącznie autorom, recenzentom, członkom RN i KR oraz wydawcy.

5. Zmiany dokonane w tekście na etapie przygotowania artykułu do publikacji nie mogą naruszać zasadniczej myśli autorów. Wszelkie modyfikacje o charakterze merytorycznym są z nimi konsultowane.
6. W przypadku podjęcia decyzji o niepublikowaniu artykułu nie może on zostać w żaden sposób wykorzystany przez wydawcę lub uczestników procesu wydawniczego bez pisemnej zgody autorów. Autorzy mogą się odwołać od decyzji o niepublikowaniu artykułu. W tym celu powinni się skontaktować z redaktorem naczelnym lub sekretarzem redakcji „WS" i przedstawić stosowną argumentację. Odwołania autorów są rozpatrywane przez redaktora naczelnego.
7. Członkowie RN i KR ani pracownicy WCN nie mogą pozostawać w jakimkolwiek konflikcie interesów w odniesieniu do artykułów zgłaszanych do publikacji. Przez konflikt interesów należy rozumieć sytuację, w której jakiekolwiek interesy lub zależności (służbowe, finansowe lub inne) mogą mieć wpływ na ocenę artykułu lub decyzję o jego publikacji.
8. W celu przeciwdziałania nierzetelności naukowej wymagane jest złożenie przez autorów oświadczenia, w którym deklarują, że zgłaszany artykuł nie narusza praw autorskich osób trzecich, nie był dotychczas publikowany i jest ich oryginalnym dziełem, a także określają swój wkład w opracowanie artykułu.
9. W celu zapewnienia wysokiej jakości recenzji wymagane jest złożenie przez recenzentów oświadczenia o przestrzeganiu zasad etyki recenzowania COPE i niewystępowaniu konfliktu interesów.
10. W przypadku uzasadnionego podejrzenia na jakimkolwiek etapie procesu wydawniczego, że autorzy dopuścili się nierzetelności naukowej (zob. pkt 3.1. Odpowiedzialność autorów), zespół redakcyjny skrupulatnie zbada sprawę ewentualnego nadużycia. Jeśli nierzetelność autorów zostanie udowodniona, to zgłoszony przez nich artykuł zostanie odrzucony przez KR, a autorzy otrzymają informację o podjętej decyzji wraz z jej uzasadnieniem.
11. Czytelnicy, którzy mają wobec autorów opublikowanego artykułu uzasadnione podejrzenia o nierzetelność naukową, powinni powiadomić o tym redaktora naczelnego lub sekretarza redakcji. Po zbadaniu sprawy ewentualnego nadużycia czytelnicy zostaną poinformowani o rezultacie przeprowadzonego postępowania. W przypadku potwierdzenia nadużycia, na łamach czasopisma zostanie zamieszczona stosowna informacja.

3.3. Odpowiedzialność recenzentów

1. Recenzenci przyjmują artykuł do recenzji tylko wtedy, gdy uznają, że:
   - posiadają odpowiednią wiedzę w określonej dziedzinie, aby rzetelnie ocenić pracę;
   - zgodnie z ich stanem wiedzy nie istnieje konflikt interesów w odniesieniu do autorów, przedstawionych w artykule badań i instytucji je finansujących, co potwierdzają w oświadczeniu;
   - mogą wywiązać się z terminu ustalonego przez redakcję, aby nie opóźniać publikacji.

2. Recenzenci są zobligowani do zachowania obiektywności i poufności oraz powstrzymania się od osobistej krytyki. Zawsze powinni uzasadnić swoją ocenę, przedstawiając stosowną argumentację.
3. Recenzenci powinni wskazać ważne dla wyników badań opublikowane prace, które w ich ocenie powinny zostać przywołane w ocenianym artykule.
4. W razie stwierdzenia wysokiego poziomu zbieżności treści recenzowanej pracy z innymi opublikowanymi materiałami lub podejrzenia innych przejawów nierzetelności naukowej recenzenci są zobowiązani poinformować o tym redakcję.
5. Po ukończeniu recenzji przechowywanie przesłanych przez redakcję materiałów (w jakiejkolwiek formie) oraz posługiwanie się nimi przez recenzentów jest niedozwolone.

3.4. Odpowiedzialność wydawcy

1. Materiały opublikowane w „WS" są chronione prawem autorskim.
2. Wydawca udostępnia pełną treść wszystkich artykułów w internecie w trybie otwartego dostępu, tj. bezpłatnie i bez technicznych ograniczeń, od 1 stycznia 2022 r. na licencji Creative Commons Uznanie autorstwa – Na tych samych warunkach 4.0 (CC BY-SA 4.0). W przypadku artykułów zgłoszonych do „WS" od 2022 r. dozwolone jest dzielenie się artykułem (kopiowanie i rozpowszechnianie go w dowolnym medium i formacie) oraz adaptowanie go (w dowolnym celu, także komercyjnym) na warunkach określonych w tej licencji. Z pozostałych artykułów zamieszczonych w czasopiśmie można korzystać w ramach otwartego dostępu, zgodnie z ustawą o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego.
3. Wydawca deklaruje gotowość do opublikowania poprawek, wyjaśnień oraz przeprosin.

4. Wymogi redakcyjne

Zgodnie z wymogami czasopisma omawiany w artykule problem badawczy powinien być jednoznacznie zdefiniowany oraz istotny dla oceny zjawisk społecznych lub gospodarczych. Artykuł powinien zawierać wyraźnie określony cel badania, precyzyjny opis badanych zjawisk i stosowanych metod, uzyskane wyniki przeprowadzonej analizy oraz autorskie wnioski.

4.1. Struktura i zawartość artykułu

Wymagane elementy artykułu:
1. Tytuł.
2. Dane autora: imię i nazwisko, afiliacja w języku polskim i angielskim, ORCID, adres e-mail. Wśród autorów artykułu wieloautorskiego należy wskazać autora korespondencyjnego.
3. Streszczenie (zalecana objętość – do 1200 znaków ze spacjami, forma bezosobowa). W przypadku artykułu opisującego badanie empiryczne powinno zawierać: cel, przedmiot, okres i metodę badania, źródła danych i najważniejsze wnioski z badania. W przypadku artykułów o innym charakterze należy podać co najmniej cel pracy, jej przedmiot i najważniejsze wnioski.

Streszczenie to podstawowe źródło informacji o artykule, warunkujące też decyzję czytelnika o zapoznaniu się z całą pracą. Dlatego powinno być przygotowane ze szczególną starannością i dbałością o umieszczenie w nim wszystkich wymaganych elementów.

4. Słowa kluczowe – najistotniejsze pojęcia lub wyrażenia użyte w pracy (nie mniej niż trzy). Powinny być zawarte w streszczeniu i/lub tytule.
5. Kod/kody z klasyfikacji Journal of Economic Literature (JEL).
6. Tłumaczenie tytułu, streszczenia i słów kluczowych (na język angielski w przypadku artykułu napisanego w języku polskim, a na język polski w przypadku artykułu napisanego w języku angielskim).
7. W artykule opisującym badanie empiryczne wymagane są następujące części:
   - wprowadzenie, zawierające syntetyczne przedstawienie zagadnień teoretycznych, uzasadnienie podjęcia danego problemu badawczego, cel badania i krytyczne odniesienie do literatury przedmiotu. W wyjątkowych przypadkach, kiedy istotne dla podjętego tematu jest obszerniejsze przedstawienie dyskusji toczącej się w literaturze, przegląd literatury może stanowić odrębną część artykułu;
   - metoda badania, uwzględniająca przedmiot i okres badania, źródła danych i zastosowane metody badawcze, w tym uzasadnienie ich wyboru;
   - wyniki badania – analiza danych oraz interpretacja wyników i odniesienie ich do rezultatów wcześniejszych badań (dyskusja). W uzasadnionych przypadkach dyskusja może stanowić odrębną część artykułu;
   - podsumowanie, które powinno być zwięzłe i odzwierciedlać istotę problemu badawczego przedstawionego w artykule, bez podawania danych liczbowych; końcowe wnioski powinny odnosić się do treści artykułu, a w szczególności do celu badania.

   Wszystkie części powinny być opatrzone numerami.
8. Bibliografia, zawierająca pełny wykaz prac i materiałów przywołanych w artykule, przygotowana zgodnie z wymogami czasopisma.

4.2. Przygotowanie artykułu

1. Artykuł powinien być utrzymany w formie bezosobowej.
2. Tekst należy zapisać alfabetem łacińskim. Nazwy własne, tytuły itp. oryginalnie zapisane innym alfabetem powinny być poddane transliteracji.
3. Nie należy stosować stylów; formatowanie należy ograniczyć do wymogów redakcyjnych.
4. Objętość artykułu łącznie ze streszczeniem, słowami kluczowymi, bibliografią, tablicami, wykresami i innymi materiałami graficznymi nie powinna być mniejsza niż 10 stron maszynopisu ani przekraczać 20 stron.
5. Edytor tekstu: Microsoft Word, format *.doc lub *.docx.
6. Krój czcionki:
   - Arial – tytuł, autor, streszczenia, słowa kluczowe, kody JEL, śródtytuły, elementy graficzne (tablice, zestawienia, wykresy, schematy), przypisy;
   - Times New Roman – tekst główny, bibliografia.
7. Wielkość czcionki:
   - 14 pkt – tytuł, autor, tytuły rozdziałów;
   - 12 pkt – tekst główny, tytuły podrozdziałów;
   - 10 pkt – pozostałe elementy.
8. Marginesy – 2,5 cm z każdej strony.

9. Interlinia – 1,5 wiersza; tablice i przypisy – 1 wiersz; przed tytułami rozdziałów i podrozdziałów oraz po nich – pusty wiersz.
10. Wcięcie akapitowe – 0,4 cm; bibliografia – bez wcięcia, wysunięcie 0,4 cm.
11. Przy wyliczeniach należy posłużyć się listą punktowaną z punktorami w postaci kropek (wysunięcie 0,4 cm, wcięcie 0 cm); wiersze (oprócz ostatniego) zakończone średnikiem.
12. Strony ponumerowane automatycznie.
13. Tablice i elementy graficzne (wykresy, mapy, schematy) muszą być przywołane w tekście.
14. Wykresy, mapy i schematy należy zamieścić w tekście głównym. Wykresy powinny być edytowalne (optymalnie wykonane w programie Excel; w przypadku wykonania w programie graficznym powinny mieć postać wektorową). Wykresy i inne materiały graficzne należy przekazać osobno w pliku programu Excel lub innym edytowalnym w pakiecie Microsoft Office. W przypadku wykonania ich w innym programie odpowiedni format pliku będzie ustalany indywidualnie.
15. Tablice muszą być edytowalne. Nie należy stosować rastrów, cieniowania, pogrubiania czy też podwójnych linii itp.
16. Wskazówki dotyczące opracowywania map znajdują się w publikacji *Mapy statystyczne. Opracowanie i prezentacja danych*, dostępnej na stronie internetowej GUS.
17. Pod tablicami i każdym elementem graficznym należy podać źródło opracowania, a także objaśnić użyte w nich skróty i symbole.
18. Literowe symbole liczb i innych wielkości niezłożonych należy zapisywać małą lub dużą literą i pismem pochyłym (np. a, A, $y(x)$, a_i); wektorów – pismem pochyłym i pogrubionym (np. $\boldsymbol{a}$, $\boldsymbol{A}$, $\boldsymbol{w}$, $\boldsymbol{y}(x)$, $\boldsymbol{w}_i$); macierzy – pismem prostym i pogrubionym (np. $\mathbf{A}$, $\mathbf{a}$, $\mathbf{M}$, $\mathbf{Y}(x)$, $\mathbf{M}_i$).
19. Objaśnienia znaków umownych w tablicach: kreska (–) – zjawisko nie wystąpiło; zero (0) – zjawisko istniało w wielkości mniejszej od 0,5; (0,0) – zjawisko istniało w wielkości mniejszej od 0,05; kropka (.) – brak informacji, konieczność zachowania tajemnicy statystycznej, wypełnienie pozycji jest niemożliwe lub niecelowe; „w tym" – oznacza, że nie podaje się wszystkich składników sumy.
20. Stosowane są następujące skróty: tablica – tabl., wykres – wykr.
21. Wszystkie zawarte w artykule informacje, dane i stwierdzenia wykraczające poza wiedzę powszechną – np. wyniki badań innych autorów, zarówno o charakterze empirycznym, jak i koncepcyjnym – muszą być opatrzone przypisem bibliograficznym. Przez wiedzę powszechną należy rozumieć informacje ogólnie znane i niebudzące wątpliwości ani kontrowersji w danej grupie społecznej, np. utworzenie GUS w 1918 r. lub powstanie UE w 1993 r. na podstawie traktatu z Maastricht. Natomiast dane statystyczne udostępniane lub publikowane np. przez GUS lub Eurostat nie należą do takich informacji. Charakteru wiedzy powszechnej nie mają również stwierdzenia odnoszące się do idei, zjawisk i procesów społecznych, politycznych czy gospodarczych. Nawet pozornie zdroworozsądkowe idee zmieniają bowiem swój sens w zależności od kultury, języka lub dyscypliny naukowej, a także bywają w rozmaity sposób konceptualizowane, jak np. pojęcie poznania w naukach społecznych.

    Podanie źródła jest konieczne niezależnie od tego, czy informacje lub stwierdzenia są ujęte w ramy cytatu, czy przedstawione bez dosłownego przytoczenia, np. w formie parafrazy. Jeżeli stwierdzenie może budzić jakiekolwiek wątpliwości odbiorców, autor powinien wskazać stosowne źródło podawanej informacji.

22. Przypisy rzeczowe, słownikowe lub informacyjne należy umieszczać na dole strony. Przypisy bibliograficzne, zgodnie ze standardem APA (American Psychological Association), należy podawać w tekście głównym.
23. Bibliografię należy przygotować zgodnie ze standardem APA.

4.3. Zasady przywoływania publikacji w treści artykułu

Wyszczególnienie	Przykład przywołania	
	w odsyłaczu	w treści zdania
Autor indywidualny		
Jeden autor	(Iksiński, 2001)	Iksiński (2001)
Dwóch autorów	(Iksiński i Nowak, 1999)	Iksiński i Nowak (1999)
Trzech autorów lub więcej	(Jankiewicz i in., 2003)	Jankiewicz i in. (2003)
Autor instytucjonalny		
Nazwa funkcjonuje jako powszechnie znany skrótowiec: pierwsze przywołanie w tekście	(International Labour Organization [ILO], 2020)	International Labour Organization (ILO, 2020)
kolejne przywołanie	(ILO, 2020)	ILO (2020)
Pełna nazwa	(Stanford University, 1995)	Stanford University (1995)
Typ publikacji		
Publikacja bez ustalonego autorstwa	(*Skrócony tytuł* …, 2015)	*Pełny tytuł* (2015)
Publikacja bez roku wydania	(Iksiński, b.r.)	Iksiński (b.r.)
Akt prawny	(Pełny tytuł)	Pełny tytuł
Strona internetowa / Zbiór danych: znana data publikacji	(Iksiński, 2020) / (Nazwa instytucji, 2020)	Iksiński (2020) / Nazwa instytucji (2020)
nieznana data publikacji	(Iksiński, b.r.) / (Nazwa instytucji, b.r.)	Iksiński (b.r.) / Nazwa instytucji (b.r.)
Rodzaj przywołania		
Przywoływanie kilku prac (porządek prac – chronologiczny, porządek autorów – alfabetyczny)	(Iksiński, 1997, 1999, 2004a, 2004b; Nowak, 2002)	Iksiński (1997, 1999, 2004a, 2004b) i Nowak (2002)
Przywoływanie publikacji za innym autorem (uwaga: w bibliografii należy wymienić tylko pracę czytaną)	(Nowakowski, 1990, za: Zieniecka, 2007)	Nowakowski (1990, za: Zieniecka, 2007)

Źródło: opracowanie na podstawie: American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th edition). https://doi.org/10.1037/0000165-000.

4.4. Przykłady opisu bibliograficznego

Bibliografia powinna być zamieszczona na końcu opracowania. Prace należy uszeregować alfabetycznie według nazwiska pierwszego autora. W przypadku dwóch lub więcej prac tego samego autora / tych samych autorów trzeba je uporządkować chronologicznie według roku publikacji. Jeśli kilka prac tego samego autora / tych samych autorów zostało opublikowanych w tym samym roku, należy podać je w kolejności alfabetycznej według tytułu i odpowiednio oznaczyć literami a, b, c itd.

Typ publikacji	Przykład opisu bibliograficznego
Artykuł w czasopiśmie	
W wersji drukowanej	Nazwisko, X. (rok). Tytuł artykułu. *Tytuł czasopisma*, *rocznik* (zeszyt), strona początku–strona końca.
Dostępny w internecie, z DOI	Nazwisko, X., Nazwisko 2, Y. (rok). Tytuł artykułu. *Tytuł czasopisma*, *rocznik*(zeszyt), strona początku–strona końca. https://doi.org/xxx.
Dostępny w internecie, bez DOI	Nazwisko, X., Nazwisko 2, Y., Nazwisko 3, Z. (rok). Tytuł artykułu. *Tytuł czasopisma*, *rocznik*(zeszyt), strona początku–strona końca. https://xxx.
Maszynopis	
Niepublikowany / przygotowywany przez autora / zgłoszony do publikacji, ale jeszcze niezaakceptowany	Nazwisko, X. (rok). *Tytuł* [maszynopis niepublikowany / w przygotowaniu / zgłoszony do publikacji].
Zaakceptowany do publikacji	Nazwisko, X. (w druku). Tytuł artykułu. *Tytuł czasopisma*.
Opublikowany nieformalnie (np. na stronie internetowej autora)	Nazwisko, X., Nazwisko 2, Y. (rok). *Tytuł artykułu*. https://xxx.
Opublikowany w trybie early view / ahead of print / online first	Nazwisko, X. (rok). Tytuł artykułu. *Tytuł czasopisma.* Early view / Ahead of print / Online first. https://xxx.
Książka	
W wersji drukowanej	Nazwisko, X. (rok). *Tytuł książki*. Wydawnictwo.
Dostępna w internecie, z DOI	Nazwisko, X. (rok). *Tytuł książki*. Wydawnictwo. https://doi.org/xxx.
Dostępna w internecie, bez DOI	Nazwisko, X. (rok). *Tytuł książki*. Wydawnictwo. https://xxx.
W przekładzie	Nazwisko, X. (rok). *Tytuł książki* (tłum. Y. Nazwisko). Wydawnictwo.
Wydanie wielotomowe: tom zatytułowany	Nazwisko, X. (rok). *Tytuł książki: nr tomu. Tytuł tomu*. Wydawnictwo.
tom niezatytułowany	Nazwisko, X. (rok). *Tytuł książki* (nr tomu). Wydawnictwo.
Kolejne wydanie	Nazwisko, X. (rok). *Tytuł książki* (nr wydania). Wydawnictwo.
Pod redakcją: w języku polskim	Nazwisko, X. (red.). (rok). *Tytuł książki*. Wydawnictwo.
w języku angielskim	Nazwisko, X. (Ed.). (rok). *Tytuł książki*. Wydawnictwo.
Rozdział w pracy zbiorowej	Nazwisko, X. (rok). Tytuł rozdziału. W: Y. Nazwisko, Z. Nazwisko 2 (red.), *Tytuł książki* (s. strona początku–strona końca). Wydawnictwo. https://doi.org/xxx lub https://xxx.
Inne prace	
Raport: autor indywidualny	Nazwisko, X. (rok). *Tytuł raportu*. Wydawnictwo.
autor instytucjonalny	Nazwa instytucji. (rok). *Tytuł raportu*. Wydawnictwo.
Working Papers	Nazwisko, X. (rok). *Tytuł pracy* (nazwa serii i numer). https://doi.org/xxx lub https://xxx.
Sesja konferencyjna / prezentacja / referat	Nazwisko, X. (rok, dzień i miesiąc). *Tytuł pracy* [typ wystąpienia, np. referat]. Nazwa konferencji, miejsce konferencji.
Rozprawa doktorska: nieopublikowana	Nazwisko, X. (rok). *Tytuł pracy* [niepublikowana rozprawa doktorska]. Nazwa instytucji nadającej tytuł doktorski.
opublikowana	Nazwisko, X. (rok). *Tytuł pracy* [rozprawa doktorska, nazwa instytucji nadającej tytuł doktorski]. https://xxx.
Akt prawny	Pełny tytuł aktu prawnego wraz z datą publikacji w dzienniku urzędowym.

Typ publikacji	Przykład opisu bibliograficznego
Strona internetowa	
Znana data publikacji, zawartość strony się nie zmienia	Nazwisko, X. (rok, dzień i miesiąc). *Tytuł*. https://xxx.
Nieznana data publikacji, zawartość strony się zmienia	Nazwa instytucji. (b.r.). *Tytuł*. Pobrane dzień, miesiąc i rok pobrania z https://xxx.
Zbiór danych	
Surowe dane nieopublikowane	Nazwisko, X. (rok wydania pracy, w której dane są wykorzystywane) [opis danych, np. surowe dane nieopublikowane dotyczące…]. Źródło danych (np. nazwa uniwersytetu).
Dane opublikowane: znana data publikacji, zawartość zbioru się nie zmienia	Nazwisko, X. (rok). *Nazwa zbioru danych* [zbiór danych]. Wydawca. https://xxx.
nieznana data publikacji, zawartość zbioru się zmienia	Nazwa instytucji. (b.r.). *Nazwa zbioru danych* [zbiór danych]. Wydawca. Pobrane dzień, miesiąc i rok pobrania z https://xxx.

Źródło: opracowanie na podstawie: American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th edition). https://doi.org/10.1037/0000165-000.

Praca przygotowana w sposób niezgodny z powyższymi wskazówkami będzie odesłana do autora z prośbą o dostosowanie formy artykułu do wymogów redakcyjnych.

ZAKRES TEMATYCZNY DZIAŁÓW
THEMATIC SCOPE OF SECTIONS

(for the English translation of the information given below, please visit ws.stat.gov.pl/AimScope)

Studia metodologiczne

W tym dziale zamieszczane są artykuły naukowe przedstawiające teoretyczne rozwiązania metodologiczne ze wskazaniem ich praktycznej użyteczności, w tym prace przeglądowe i porównawcze oraz dotyczące etyki w statystyce. Poruszane w nich zagadnienia obejmują różne dziedziny statystyki, ekonomii matematycznej i ekonometrii. Omawiane rezultaty badawcze mogą znaleźć efektywne zastosowanie w badaniach empirycznych oraz analizach statystycznych i służyć podnoszeniu ich jakości, jak również powiększeniu zasobu informacyjnego.

Statystyka w praktyce

Dział ten zawiera artykuły poświęcone nowatorskim zastosowaniom w praktyce znanych narzędzi i modeli statystycznych oraz analizie i ocenie statystycznej zjawisk społeczno-ekonomicznych i innych; zamieszczane tu prace opierają się w szczególności na danych pochodzących z zasobów statystyki publicznej. Zastosowania w praktyce obejmują również wykorzystanie narzędzi informatycznych do uzyskiwania i przetwarzania informacji statystycznych, naliczania danych wynikowych, ich prezentacji i rozpowszechniania. Może to też dotyczyć opracowań stosujących nowoczesne techniki programistyczne pozwalające na efektywną komunikację z systemami informacyjnymi oraz ułatwiające wykorzystanie danych wynikowych. Publikowane są także artykuły sygnalizujące problemy związane z projektowaniem badań statystycznych, uzyskiwaniem, integracją i przetwarzaniem danych oraz generowaniem wynikowych informacji statystycznych i kontrolą ich ujawniania wraz z propozycjami efektywnych rozwiązań w tym zakresie.

Studia interdyscyplinarne. Wyzwania badawcze

To blok tematyczny zawierający artykuły wskazujące i podejmujące wyzwania badawcze, które są szczególnie istotne ze względu na rosnące potrzeby współczesnych użytkowników danych statystycznych i wymagają zaangażowania znacznych nakładów pracy, środków oraz rozwiązań z różnych dziedzin nauki i techniki. W dziale tym publikowane są również opracowania dotyczące: wykorzystania technologii informacyjnych i komunikacyjnych (ICT), gospodarki opartej na wiedzy, problematyki innowacyjności, przepływu informacji we współczesnym społeczeństwie oraz przetwarzania i analizy zagadnień związanych z data science i big data, a zatem problematyki bardzo często powiązanej z działaniami interdyscyplinarnymi.

Edukacja statystyczna

W tym dziale zamieszczane są artykuły dotyczące metod i efektów nauczania statystyki oraz popularyzacji myślenia statystycznego. Odnosi się to zwłaszcza do problemów związanych z kształceniem w zakresie umiejętności stosowania statystyki na wszystkich poziomach edukacji, a także do wykorzystywania nowoczesnych koncepcji i metod dydaktycznych oraz pomocy naukowych w nauczaniu statystyki. Uwaga skoncentrowana jest na rozumieniu prawdopodobieństwa i statystyki, badaniach z zakresu nauczania statystyki, postaw i zachowań społecznych w odniesieniu do tej dziedziny wiedzy, jak również na rozumieniu informacji statystycznych. Ponadto ukazywane są problemy związane z prezentacją danych statystycznych oraz ich interpretacją w powszechnym obiegu informacyjnym, np. w środkach społecznego przekazu.

Z dziejów statystyki

Prace publikowane w tym dziale poświęcone są historii prowadzenia obserwacji statystycznych oraz rozwoju ich metodologii i narzędzi. Ponadto zamieszczane są tu informacje dotyczące życia i osiągnięć zawodowych wybitnych statystyków, jak również najważniejszych instytucji i organizacji statystycznych w Polsce i za granicą.

Dyskusje. Recenzje. Informacje

Jedyny dział zawierający teksty nierecenzowane i niemające charakteru artykułów naukowych. Obejmuje informacje o najważniejszych wydarzeniach dotyczących statystyki polskiej i międzynarodowej, a także sprawozdania z konferencji naukowych, recenzje książek i opracowań z zakresu statystyki i jej zastosowań, rekomendacje nowych, istotnych i ciekawych pozycji wydawniczych z tego obszaru wiedzy, jak również odpowiedzi autorów na recenzje oraz polemiki, dyskusje i sprostowania dotyczące artykułów zamieszczonych na łamach czasopisma.